PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classificati n 6 : C12N 15/31, 15/63, 15/00, C12P 21/02	A1	**(11) International Publication Number:** WO 96/33276 **(43) International Publication Date:** 24 October 1996 (24.10.96)

(21) International Application Number: PCT/US96/05320

(22) International Filing Date: 22 April 1996 (22.04.96)

(30) Priority Data:

08/426,787	21 April 1995 (21.04.95)	US
08/476,102	7 June 1995 (07.06.95)	US
08/487,429	7 June 1995 (07.06.95)	US

(71) Applicants: HUMAN GENOME SCIENCES, INC. [US/US]; 9410 Key West Avenue, Rockville, MD 20850 (US). JOHNS HOPKINS UNIVERSITY [US/US]; 720 Rutland Avenue, Baltimore, MD 21205 (US).

(72) Inv ntors: FLEISCHMANN, Robert, D.; 470 Tschiffely Square Road, Gaithersburg, MD 20878 (US). ADAMS, Mark, D.; 15205 Dufief Drive, N. Potomac, MD 20878 (US). WHITE, Owen; Apartment #202, 886 Quince Orchard Boulevard, Gaithersburg, MD 20878 (US). SMITH, Hamilton, O.; 8222 Carrbridge Circle, Towson, MD 21204 (US). VENTER, J., Craig; 11915 Glen Mill Road, Potomac, MD 20854 (US).

(74) Agents: GOLDSTEIN, Jorge, A. et al.; Sterne, Kessler, Goldstein & Fox P.L.L.C., Suite 600, 1100 New York Avenue, Washington, DC 20005-3934 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: NUCLEOTIDE SEQUENCE OF THE HAEMOPHILUS INFLUENZAE Rd GENOME, FRAGMENTS THEREOF, AND USES THEREOF

(57) Abstract

The present invention provides the sequencing of the entire genome of *Haemophilus influenzae* Rd, SEQ ID NO:1. The present invention further provides the sequence information stored on computer readable media, and computer-based systems and methods which facilitate its use. In addition to the entire genomic sequence, the present invention identifies over 1700 protein encoding fragments of the genome and identifies, by position relative to a unique *Not I* restriction endonuclease site, any regulatory elements which modulate the expression of the protein encoding fragments of the *Haemophilus* genome.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia
AT	Austria
AU	Australia
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CS	Czechoslovakia
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgystan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakhstan
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam

Nucleotide Sequence of the *Haemophilus influenzae* Rd Genome, Fragments Thereof, and Uses Thereof

Part of the work performed during development of this invention utilized U.S. Government funds. The government may have certain rights in this invention. NIH-5R01GM48251

Field of the Invention

The present invention relates to the field of molecular biology. The present invention discloses compositions comprising the nucleotide sequence of *Haemophilus influenzae*, fragments thereof and usage in industrial fermentation and pharmaceutical development.

Background of the Invention

The complete genome sequence from a free living cellular organism has never been determined. The first mycobacterium sequence should be completed by 1996, while *E. coli* and *S. cerevisae* are expected to be completed before 1998. These are being done by random and/or directed sequencing of overlapping cosmid clones. No one has attempted to determine sequences of the order of a megabase or more by a random shotgun approach.

H. influenzae is a small (approximately 0.4 x 1 micron) non-motile, non-spore forming, germ-negative bacterium whose only natural host is human. It is a resident of the upper respiratory mucosa of children and adults and causes otitis media and respiratory tract infections mostly in children. The most serious complication is meningitis, which produces neurological sequelae

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO 9633276A1 I >

in up to 50% of affected children. Six *H. influenzae* serotypes (a through f) have been identified based on immunologically distinct capsular polysaccharide antigens. A number of non-typeable strains are also known. Serotype b accounts for the majority of human disease.

Interest in the medically important aspects of *H. influenzae* biology has focused particularly on those genes which determine virulence characteristics of the organism. A number of the genes responsible for the capsular polysaccharide have been mapped and sequenced (Kroll *et al.*, *Mol. Microbiol.* 5(6):1549-1560 (1991)). Several outer membrane protein (OMP) genes have been identified and sequenced (Langford *et al.*, *J. Gen. Microbiol. 138*:155-159 (1992)). The lipoligosaccharide (LOS) component of the outer membrane and the genes of its synthetic pathway are under intensive study (Weiser *et al.*, *J. Bacteriol. 172*:3304-3309 (1990)). While a vaccine has been available since 1984, the study of outer membrane components is motivated to some extent by the need for improved vaccines. Recently, the catalase gene was characterized and sequenced as a possible virulence-related gene (Bishni *et al.*, in press). Elucidation of the *H. influenzae* genome will enhance the understanding of how *H. influenzae* causes invasive disease and how best to combat infection.

H. influenzae possesses a highly efficient natural DNA transformation system which has been intensively studied in the non-encapsulated (R), serotype d strain (Kahn and Smith, *J. Membrane Biology 81*:89-103 (1984)). At least 16 transformation-specific genes have been identified and sequenced. Of these, four are regulatory (Redfield, *J. Bacteriol. 173*:5612-5618 (1991), and Chandler, *Proc. Natl. Acad. Sci. USA 89*:1626-1630 (1992)), at least two are involved in recombination processes (Barouki and Smith, *J. Bacteriol. 163*(2):629-634 (1985)), and at least seven are targeted to the membranes and periplasmic space (Tomb *et al.*, *Gene 104*:1-10 (1991), and Tomb, *Proc. Natl. Acad. Sci. USA 89*:10252-10256 (1992)), where they appear to function as structural components or in the assembly of the DNA transport machinery. *H. influenzae* Rd transformation shows a number of interesting features including

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO__9633276A1_I_>

sequence-specific DNA uptake, rapid uptake of several double-stranded DNA molecules per competent cell into a membrane compartment called the transformasome, linear translocation of a single strand of the donor DNA into the cytoplasm, and synapsis and recombination of the strand with the chromosome by a single-strand displacement mechanism. The *H. influenzae* Rd transformation system is the most thoroughly studied of the gram-negative systems and distinct in a number of ways from the gram-positive systems.

The size of *H. influenzae* Rd genome has been determined by pulsed-field agarose gel electrophoresis of restriction digests to be approximately 1.9 Mb, making its genome approximately 40% the size of *E. coli* (Lee and Smith, *J. Bacteriol. 170*:4402-4405 (1988)). The restriction map of *H. influenzae* is circular (Lee *et al.*, *J. Bacteriol. 171*:3016-3024 (1989), and Redfield and Lee, "*Haemophilus influenzae* Rd", pp. 2110-2112, In O'Brien, S.J. (ed), Genetic Maps: Locus Maps of Complex Genomes, Cold Spring Harbor Press, New York). Various genes have been mapped to restriction fragments by Southern hybridization probing of restriction digest DNA bands. This map will be valuable in verification of the assembly of a complete genome sequence from randomly sequenced fragments. GenBank currently contains about 100 kb of non-redundant *H. influenzae* DNA sequences. About half are from serotype b and half from Rd.

Summary of the Invention

The present invention is based on the sequencing of the *Haemophilus influenzae* Rd genome. The primary nucleotide sequence which was generated is provided in SEQ ID NO:1.

The present invention provides the generated nucleotide sequence of the *Haemophilus influenzae* Rd genome, or a representative fragment thereof, in a form which can be readily used, analyzed, and interpreted by a skilled artisan. In one embodiment, present invention is provided as a contiguous

BNSDOCID: <WO__9633276A1 I_>

string of primary sequence inf rmation corresponding to the nucleotide sequence depicted in SEQ ID NO:1.

The present invention further provides nucleotide sequences which are at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1.

The nucleotide sequence of SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence which is at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 may be provided in a variety of mediums to facilitate its use. In one application of this embodiment, the sequences of the present invention are recorded on computer readable media. Such media includes, but is not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media.

The present invention further provides systems, particularly computer-based systems which contain the sequence information herein described stored in a data storage means. Such systems are designed to identify commercially important fragments of the *Haemophilus influenzae* Rd genome.

Another embodiment of the present invention is directed to isolated fragments of the *Haemophilus influenzae* Rd genome. The fragments of the *Haemophilus influenzae* Rd genome of the present invention include, but are not limited to, fragments which encode peptides, hereinafter open reading frames (ORFs), fragments which modulate the expression of an operably linked ORF, hereinafter expression modulating fragments (EMFs), fragments which mediate the uptake of a linked DNA fragment into a cell, hereinafter uptake modulating fragments (UMFs), and fragments which can be used to diagnose the presence of *Haemophilus influenzae* Rd in a sample, hereinafter, diagnostic fragments (DFs).

Each of the ORF fragments of the *Haemophilus influenzae* Rd genome disclosed in Tables 1(a) and 2, and the EMF found 5' to the ORF, can be used in numerous ways as polynucleotide reagents. The sequences can be used as diagnostic probes or diagn stic amplification primers for the presence of a

specific microbe in a sample, for the production of commercially important pharmaceutical agents, and to selectively control gene expressi n.

The present invention further includes recombinant constructs comprising one or more fragments of the *Haemophilus influenzae* Rd genome of the present invention. The recombinant constructs of the present invention comprise vectors, such as a plasmid or viral vector, into which a fragment of the *Haemophilus influenzae* Rd has been inserted.

The present invention further provides host cells containing any one of the isolated fragments of the *Haemophilus influenzae* Rd genome of the present invention. The host cells can be a higher eukaryotic host such as a mammalian cell, a lower eukaryotic cell such as a yeast cell, or can be a procaryotic cell such as a bacterial cell.

The present invention is further directed to isolated proteins encoded by the ORFs of the present invention. A variety of methodologies known in the art can be utilized to obtain any one of the proteins of the present invention. At the simplest level, the amino acid sequence can be synthesized using commercially available peptide synthesizers. In an alternative method, the protein is purified from bacterial cells which naturally produce the protein. Lastly, the proteins of the present invention can alternatively be purified from cells which have been altered to express the desired protein.

The invention further provides methods of obtaining homologs of the fragments of the *Haemophilus influenzae* Rd genome of the present invention and homologs of the proteins encoded by the ORFs of the present invention. Specifically, by using the nucleotide and amino acid sequences disclosed herein as a probe or as primers, and techniques such as PCR cloning and colony/plaque hybridization, one skilled in the art can obtain homologs.

The invention further provides antibodies which selectively bind one of the proteins of the present invention. Such antibodies include both monoclonal and polyclonal antibodies.

SUBSTITUTE SHEET (RULE 26)

The invention further provides hybridomas which produce the above-described antibodies. A hybridoma is an immortalized cell line which is capable of secreting a specific monoclonal antibody.

The present invention further provides methods of identifying test samples derived from cells which express one of the ORF of the present invention, or homolog thereof. Such methods comprise incubating a test sample with one or more of the antibodies of the present invention, or one or more of the DFs of the present invention, under conditions which allow a skilled artisan to determine if the sample contains the ORF or product produced therefrom.

In another embodiment of the present invention, kits are provided which contain the necessary reagents to carry out the above-described assays.

Specifically, the invention provides a compartmentalized kit to receive, in close confinement, one or more containers which comprises: (a) a first container comprising one of the antibodies, or one of the DFs of the present invention; and (b) one or more other containers comprising one or more of the following: wash reagents, reagents capable of detecting presence of bound antibodies or hybridized DFs.

Using the isolated proteins of the present invention, the present invention further provides methods of obtaining and identifying agents capable of binding to a protein encoded by one of the ORFs of the present invention. Specifically, such agents include antibodies (described above), peptides, carbohydrates, pharmaceutical agents and the like. Such methods comprise the steps of:

(a) contacting an agent with an isolated protein encoded by one of the ORFs of the present invention; and

(b) determining whether the agent binds to said protein.

The complete genomic sequence of *H. influenzae* will be of great value to all laboratories working with this organism and for a variety of commercial purposes. Many fragments of the *Haemophilus influenzae* Rd genome will be immediately identified by similarity searches against GenBank or protein

BNSDOCID: <WO___9633276A1_I_>

databases and will be of immediate value to *Haemophilus* researchers and for immediate commercial value for the production of proteins or to control gene expression. A specific example concerns PHA synthase. It has been reported that polyhydroxybutyrate is present in the membranes of *H. influenzae* Rd and that the amount correlates with the level of competence for transformation. The PHA synthase that synthesizes this polymer has been identified and sequenced in a number of bacteria, none of which are evolutionarily close to *H. influenzae*. This gene has yet to be isolated from *H. influenzae* by use of hybridization probes or PCR techniques. However, the genomic sequence of the present invention allows the identification of the gene by utilizing search means described below.

Developing the methodology and technology for elucidating the entire genomic sequence of bacterial and other small genomes has and will greatly enhance the ability to analyze and understand chromosomal organization. In particular, sequenced genomes will provide the models for developing tools for the analysis of chromosome structure and function, including the ability to identify genes within large segments of genomic DNA, the structure, position, and spacing of regulatory elements, the identification of genes with potential industrial applications, and the ability to do comparative genomic and molecular phylogeny.

Description of the Figures

Figure 1 - restriction map of the *Haemophilus influenzae* Rd genome.

Figure 2 - Block diagram of a computer system 102 that can be used to implement the computer-based systems of present invention.

Figure 3 - A comparison of experimental coverage of up to approximately 4000 random sequence fragments assembled with AutoAssembler (squares) as compared to Lander-Waterman prediction for a 2.5 Mb genome (triangles) and a 1.6 Mb gen me (circles) with a 460 bp average sequence length and a 25 bp overlap.

Figure 4 - Data flow and computer programs used to manage, assemble, edit, and ann tate the *H. influenzae* genome. Both Macintosh and Unix platforms are used to handle the AB 373 sequence data files (Kerlavage *et al., Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*, IEEE Computer Society Press, Washington D.C., 585 (1993)). **Factura** (AB) is a Macintosh program designed for automatic vector sequence removal and end trimming of sequence files. The program **esp** runs on a Macintosh platform and parses the feature data extracted from the sequence files by **Factura** to the Unix based *H. influenzae* relational database. Assembly is accomplished by retrieving a specific set of sequence files and their associated features using **stp**, an X-windows graphical interface and control program which can retrieve sequences from the *H. influenzae* database using user-defined or standard SQL queries. The sequence files were assembled using **TIGR Assembler**, an assembly engine designed at TIGR for rapid and accurate assembly of thousands of sequence fragments. **TIGR Editor** is a graphical interface which can parse the aligned sequence files from **TIGR Assembler** output and display the alignment and associated electropherograms for contig editing. Identification of putative coding regions was performed with **Genemark** (Borodovsky and McIninch, *Computers Chem. 17(2)*:123 (1993)), a Markov and Bayes modeled program for predicting gene locations, and trained on a *H. influenzae* sequence data set. Peptide searches were performed against the three reading frames of each **Genemark** predicted coding region using **blaze** (Brutlag *et al., Computers Chem. 17*:203 (1993)) run on a Maspar MP-2 massively parallel computer with 4096 microprocessors. Results from each frame were combined into a single output file by **mblzt**. Optimal protein alignments were obtained using the program **praze** which extends alignments across potential frameshifts. The output was inspected using a custom graphic viewing program, **gbyob**, that interacts directly with the *H. influenzae* database. The alignments were further used to identify potential frameshift errors and were targeted for additional editing.

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

Figure 5 - A circular representation of the *H. influenzae* Rd chromosome illustrating the location of each predicted coding region containing a database match as well as selected global features of the genome. **Outer perimeter**: The location of the unique *Not*I restriction site (designated as nucleotide 1), the *Rsr*II sites, and the *Sma*I sites. *Outer concentric circle*: The location of each identified coding region for which a gene identification was made. Each coding region location is coded as to role according to the color code in Fig. 6. **Second concentric circle**: Regions of high G/C content (> 42%, red; > 40%, blue) and high A/T content (> 66%, black; > 64%, green). High G/C content regions are specifically associated with the 6 ribosomal operons and the mu-like prophage. **Third concentric circle**: Coverage by lambda clones (blue). Over 300 lambda clones were sequenced from each end to confirm the overall structure of the genome and identify the 6 ribosomal operons. **Fourth concentric circle**: The locations of the 6 ribosomal operons (green), the tRNAs (black) and the cryptic mu-like prophage (blue). **Fifth concentric circle**: Simple tandem repeats. The locations of the following repeats are shown: CTGGCT, GTCT, ATT, AATGGC, TTGA, TTGG, TTTA, TTATC, TGAC, TCGTC, AACC, TTGC, CAAT, CCAA. The putative origin of replication is illustrated by the outward pointing arrows (green) originating near base 603,000. Two potential termination sequences are shown near the opposite midpoint of the circle (red).

Figures 6(A)-6(D)- Complete map of the *H. influenzae* Rd genome. Predicted coding regions are shown on each strand. rRNA and tRNA genes are shown as lines and triangles, respectively. Genes are color-coded by role category as described in the legend. GeneID numbers correspond to those in Tables 1(a), 1(b) and 2. Where possible, three-letter designations are also provided.

Figure 7 - A comparison f the region of the *H. influenzae* chromosome containing the 8 genes of the fimbrial gene cluster present in *H.*

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO 9633276A1 I >

influenzae type b and the same region in *H. influenzae* Rd. The region is flanked by the *pepN* and *purE* genes in both organisms. However in the non-infectious Rd strain the 8 genes of the fimbrial gene cluster have been excised. A 172 bp spacer region is located in this region in the Rd strain and continues to be flanked by the *pepN* and *purE* genes.

Figure 8 - Hydrophobicity analysis of five predicted channel-proteins. The amino acid sequences of five predicted coding regions that do not display homology with known peptide sequences (GenBank release 87), each exhibit multiple hydrophobic domains that are characteristic of channel-forming proteins. The predicted coding region sequences were analyzed by the Kyte-Doolittle algorithm (Kyte and Doolittle, *J. Mol. Biol. 157*:105 (1982)) (with a range of 11 residues) using the GeneWorks software package (Intelligenetics).

Detailed Description of the Preferred Embodiments

The present invention is based on the sequencing of the *Haemophilus influenzae* Rd genome. The primary nucleotide sequence which was generated is provided in SEQ ID NO:1. As used herein, the "primary sequence" refers to the nucleotide sequence represented by the IUPAC nomenclature system.

The sequence provided in SEQ ID NO:1 is oriented relative to a unique *Not* I restriction endonuclease site found in the *Haemophilus influenzae* Rd genome. A skilled artisan will readily recognize that this start/stop point was chosen for convenience and does not reflect a structural significance.

The present invention provides the nucleotide sequence of SEQ ID NO:1, or a representative fragment thereof, in a form which can be readily used, analyzed, and interpreted by a skilled artisan. In one embodiment, the sequence is provided as a contiguous string of primary sequence information corresponding to the nucleotide sequence provided in SEQ ID NO:1.

As used herein, a "representative fragment of the nucleotide sequence depicted in SEQ ID NO:1" refers to any portion f SEQ ID NO:1 which is not presently represented within a publicly available database. Preferred representative fragments of the present invention are *Haemophilus influenzae* open reading frames, expression modulating fragments, uptake modulating fragments, and fragments which can be used to diagnose the presence of *Haemophilus influenzae* Rd in sample. A non-limiting identification of such preferred representative fragments is provided in Tables 1(a) and and 2.

The nucleotide sequence information provided in SEQ ID NO:1 was obtained by sequencing the *Haemophilus influenzae* Rd genome using a megabase shotgun sequencing method. Using three parameters of accuracy discussed in the Examples below, the present inventors have calculated that the sequence in SEQ ID NO:1 has a maximum accuracy of 99.98%. Thus, the nucleotide sequence provided in SEQ ID NO:1 is a highly accurate, although not necessarily a 100% perfect, representation of the nucleotide sequence of the *Haemophilus influenzae* Rd genome.

As discussed in detail below, using the information provided in SEQ ID NO:1 and in Tables 1(a) and 2 together with routine cloning and sequencing methods, one of ordinary skill in the art will be able to clone and sequence all "representative fragments" of interest including open reading frames (ORFs) encoding a large variety of *Haemophilus influenzae* proteins. In very rare instances, this may reveal a nucleotide sequence error present in the nucleotide sequence disclosed in SEQ ID NO: 1. Thus, once the present invention is made available (i.e., once the information in SEQ ID NO:1 and Tables 1(a) and 2 have been made available), resolving a rare sequencing error in SEQ ID NO:1 will be well within the skill of the art. Nucleotide sequence editing software is publicly available. For example, Applied Biosystem's (AB) AutoAssembler™ can be used as an aid during visual inspection of nucleotide sequences.

SUBSTITUTE SHEET (RULE 26)

SNSDOCID: <WO 9633276A1 I >

Even if all of the very rare sequencing errors in SEQ ID NO:1 were corrected, the resulting nucleotide sequence would still be at least 99.9% identical to the nucleotide sequence in SEQ ID NO:1.

The nucleotide sequences of the genomes from different strains of *Haemophilus influenzae* differ slightly. However, the nucleotide sequence of the genomes of all *Haemophilus influenzae* strains will be at least 99.9% identical to the nucleotide sequence provided in SEQ ID NO:1.

Thus, the present invention further provides nucleotide sequences which are at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 in a form which can be readily used, analyzed and interpreted by the skilled artisan. Methods for determining whether a nucleotide sequence is at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 are routine and readily available to the skilled artisan. For example, the well known **fasta** algothrithm (Pearson and Lipman, *Proc. Natl. Acad. Sci. USA 85*:2444 (1988)) can be used to generate the percent identity of nucleotide sequences.

Computer Related Embodiments

The nucleotide sequence provided in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1 may be "provided" in a variety of mediums to facilitate use thereof. As used herein, provided refers to a manufacture, other than an isolated nucleic acid molecule, which contains a nucleotide sequence of the present invention, i.e., the nucleotide sequence provided in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1. Such a manufacture provides the *Haemophilus influenzae* Rd genome or a subset thereof (e.g., a *Haemophilus Influenzae* Rd open reading frame (ORF)) in a form which allows a skilled artisan to examine the manufacture using means not directly applicable to examining the *Haemophilus influenzae* Rd genome or a subset thereof as it exists in nature or in purified form.

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on c mputer readable media. As used herein, "computer readable media" refers to any medium which can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently know methods for recording information on computer readable medium to generate manufactures comprising the nucleotide sequence information of the present invention.

A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and MicroSoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of dataprocessor structuring formats (e.g. text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

BNSDOCID: <WO _9633276A1_I >

By providing the nucleotide sequence of SEQ ID NO: 1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1 in computer readable form, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul *et al., J. Mol. Biol. 215*:403-410 (1990)) and BLAZE (Brutlag *et al., Comp. Chem. 17*:203-207 (1993)) search algorithms on a Sybase system was used to identify open reading frames (ORFs) within the *Haemophilus influenzae* Rd genome which contain homology to ORFs or proteins from other organisms. Such ORFs are protein encoding fragments within the *Haemophilus influenzae* Rd genome and are useful in producing commercially important proteins such as enzymes used in fermentation reactions and in the production of commercially useful metabolites.

The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the *Haemophilus influenzae* Rd genome.

As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention.

As stated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means. As used herein, "data storage means" refers to memory which can store nucleotide sequence

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention.

As used herein, "search means" refers to one or more programs which are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the *Haemophilus influenzae* Rd genome which match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are and can be used in the computer-based systems of the present invention. Examples of such software includes, but is not limited to, MacPattern (EMBL), BLASTN and BLASTX (NCBIA). A skilled artisan can readily recognize that any one of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

As used herein, a "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that searches for commercially important fragments of the *Haemophilus influenzae* Rd genome, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymic active sites and signal sequences. Nucleic acid target motifs include,

SUBSTITUTE SHEET (RULE 26)

INSDOCID: <WO 9633276A1_I_>

but are not limited to, promoter sequences, hairpin structures and inducible expression elements (protein binding sequences).

A variety of structural formats for the input and output means can be used to input and output the information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the *Haemophilus influenzae* Rd genome possessing varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments of the *Haemophilus influenzae* Rd genome. In the present examples, implementing software which implement the BLAST and BLAZE algorithms (Altschul *et al.*, *J. Mol. Biol. 215*:403-410 (1990)) was used to identify open reading frames within the *Haemophilus influenzae* Rd genome. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention.

One application of this embodiment is provided in Figure 2. Figure 2 provides a block diagram of a computer system 102 that can be used to implement the present invention. The computer system 102 includes a processor 106 connected to a bus 104. Also connected to the bus 104 are a main memory 108 (preferably implemented as random access memory, RAM) and a variety of secondary storage devices 110, such as a hard drive 112 and a removable medium storage device 114. The removable medium storage device 114 may represent, for example, a floppy disk drive, a CD-ROM drive, a magnetic tape drive, etc. A removable storage medium 116 (such as a floppy disk, a compact disk, a magnetic tape, etc.) containing control logic and/ r data recorded therein may be inserted into the rem vable medium storage device 114. The computer system 102 includes appropriate software

BNSDOCID: <WO___9633276A1_I_>

f r reading the control logic and/or the data from the removable medium storage device 114 once inserted in the removable medium storage device 114.

A nucleotide sequence of the present invention may be stored in a well known manner in the main memory 108, any of the secondary storage devices 110, and/or a removable storage medium 116. Software for accessing and processing the genomic sequence (such as search tools, comparing tools, etc.) reside in main memory 108 during execution.

Biochemical Embodiments

Another embodiment of the present invention is directed to isolated fragments of the *Haemophilus influenzae* Rd genome. The fragments of the *Haemophilus influenzae* Rd genome of the present invention include, but are not limited to fragments which encode peptides, hereinafter open reading frames (ORFs), fragments which modulate the expression of an operably linked ORF, hereinafter expression modulating fragments (EMFs), fragments which mediate the uptake of a linked DNA fragment into a cell, hereinafter uptake modulating fragments (UMFs), and fragments which can be used to diagnose the presence of *Haemophilus influenzae* Rd in a sample, hereinafter diagnostic fragments (DFs).

As used herein, an "isolated nucleic acid molecule" or an "isolated fragment of the *Haemophilus influenzae* Rd genome" refers to a nucleic acid molecule possessing a specific nucleotide sequence which has been subjected to purification means to reduce, from the composition, the number of compounds which are normally associated with the composition. A variety of purification means can be used to generated the isolated fragments of the present invention. These include, but are not limited to methods which separate constituents of a solution based on charge, solubility, or size.

In one embodiment, *Haemophilus influenaze* Rd DNA can be mechanically sheared to produce fragments of 15-20 kb in length. These fragments can then be used to generate an *Haemophilus influenzae* Rd library

BNSDOCID: <WO___9633276A1_I_>

by inserting them into labda clones as described in the Examples below. Primers flanking, for examiple, an ORF provided in Table 1(a) can then be generated using nucleotide sequence information provided in SEQ ID NO:1. PCR cloning can then be used to isolate the ORF from the lambda DNA library. PCR cloning is well known in the art. Thus, given the availability of SEQ ID NO:1, Table 1(a) and Table 2, it would be routine to isolate any ORF or other nucleic acid fragment of the present invention.

The isolated nucleic acid molecules of the present invention include, but are not limited to single stranded and double stranded DNA, and single stranded RNA.

As used herein, an "open reading frame," ORF, means a series of triplets coding for amino acids without any termination codons and is a sequence translatable into protein. Tables 1a, 1b and 2 identify ORFs in the *Haemophilus influenzae* Rd genome. In particular, Table 1a indicates the location of ORFs within the *Haemophilus influenzae* genome which encode the recited protein based on homology matching with protein sequences from the organism appearing in parentheticals (see the fourth column of Table 1(a)).

The first column of Table 1(a) provides the "GeneID" of a particular ORF. This information is useful for two reasons. First, the complete map of the *Haemophilus influenzae* Rd genome provided in Figures 6(A)-6(D) refers to the ORFs according to their GeneID numbers. Second, Table 1(b) uses the GeneID numbers to indicate which ORFs were provided previously in a public database.

The second and third columns in Table 1(a) indicate an ORFs position in the nucleotide sequence provided in SEQ ID NO:1. One of ordinary skill will recognize that ORFs may be oriented in opposite directions in the *Haemophilus influenae* genome. This is reflected in columns 2 and 3.

The fifth column of Table 1(a) indicates the percent identity of the protein encoded for by an ORF to the corresponding protein from the orgaism appearing in parentheticals in the fourth column.

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

The sixth column of Table 1(a) indicates the percent similarity of the protein encoded for by an ORF to the corresponding protein from the organism appearing in parentheticals in the fourth column. The concepts of percent identity and percent similarity of two polypeptide sequences is well understood in the art. For example, two polypeptides 10 amino acids in length which differ at three amino acid positions (e.g., at positions 1, 3 and 5) are said to have a percent identity of 70%. However, the same two polypeptides would be deemed to have a percent similarity of 80% if, for example at position 5, the amino acids moieties, although not identical, were "similar" (i.e., possessed similar biochemical characteristics).

The seventh column in Table 1(a) indicates the lenth of the amino acid homology match.

Table 2 provides ORFs of the *Haemophilus influenzae* Rd genome which encode polypeptide sequences which did not elicit a "homology match" with a known protein sequence from another organism. Further details concerning the algorithms and criteria used for homology searches are provided in the Examples below.

A skilled artisan can readily identify ORFs in the *Haemophilus influenzae* Rd genome other than those listed in Tables 1(a), 1(b) and 2, such as ORFs which are overlapping or encoded by the opposite strand of an identified ORF in addition to those ascertainable using the computer-based systems of the present invention.

As used herein, an "expression modulating fragment," EMF, means a series of nucleotide molecules which modulates the expression of an operably linked ORF or EMF.

As used herein, a sequence is said to "modulate the expression of an operably linked sequence" when the expression of the sequence is altered by the presence of the EMF. EMFs include, but are not limited to, promoters, and promoter modulating sequences (inducible elements). One class of EMFs are fragments which induce the expression or an operably linked ORF in response to a specific regulatory factor or physiological event. A review of

BNSDOCID: <WO___9633276A1_I_>

known EMFs from *Haemophilus* are described by (Tomb *et al. Gene 104*:1-10 (1991), Chandler, M. S., *Proc. Natl. Acad. Sci. USA 89*:1626-1630 (1992).

EMF sequences can be identified within the *Haemophilus influenzae* Rd genome by their proximity to the ORFs provided in Tables 1(a), 1(b) and 2. An intergenic segment, or a fragment of the intergenic segment, from about 10 to 200 nucleotides in length, taken 5′ from any one of the ORFs of Tables 1(a), 1(b), or 2 will modulate the expression of an operably linked 3′ ORF in a fashion similar to that found with the naturally linked ORF sequence. As used herein, an "intergenic segment" refers to the fragments of the *Haemophilus* genome which are between two ORF(s) herein described. Alternatively, EMFs can be identified using known EMFs as a target sequence or target motif in the computer-based systems of the present invention.

The presence and activity of an EMF can be confirmed using an EMF trap vector. An EMF trap vector contains a cloning site 5′ to a marker sequence. A marker sequence encodes an identifiable phenotype, such as antibiotic resistance or a complementing nutrition auxotrophic factor, which can be identified or assayed when the EMF trap vector is placed within an appropriate host under appropriate conditions. As described above, a EMF will modulate the expression of an operably linked marker sequence. A more detailed discussion of various marker sequences is provided below.

A sequence which is suspected as being a EMF is cloned in all three reading frames in one or more restriction sites upstream from the marker sequence in the EMF trap vector. The vector is then transformed into an appropriate host using known procedures and the phenotype of the transformed host in examined under appropriate conditions. As described above, an EMF will modulate the expression of an operably linked marker sequence.

As used herein, an "uptake modulating fragment," UMF, means a series of nucleotide molecules which mediate the uptake of a linked DNA fragment into a cell. UMFs can be readily identified using known UMFs as a target sequence or target motif with the computer-based systems described above.

BNSDOCID: <WO___9633276A1_I_>

The presence and activity of a UMF can be confirmed by attaching the suspected UMF to a marker sequence. The resulting nucleic acid molecule is then incubated with an appropriate host under appropriate conditions and the uptake of the marker sequence is determined. As described above, a UMF will increase the frequency of uptake of a linked marker sequence. A review of DNA uptake in *Haemophilus* is provided by Goodgall, S.H., *et al.*, *J. Bact.* *172*:5924-5928 (1990).

As used herein, a "diagnostic fragment," DF, means a series of nucleotide molecules which selectively hybridize to *Haemophilus influenzae* sequences. DFs can be readily identified by identifying unique sequences within the *Haemophilus influenzae* Rd genome, or by generating and testing probes or amplification primers consisting of the DF sequence in an appropriate diagnostic format which determines amplification or hybridization selectivity.

The sequences falling within the scope of the present invention are not limited to the specific sequences herein described, but also include allelic and species variations thereof. Allelic and species variations can be routinely determined by comparing the sequence provided in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1 with a sequence from another isolate of the same species. Furthermore, to accommodate codon variability, the invention includes nucleic acid molecules coding for the same amino acid sequences as do the specific ORFs disclosed herein. In other words, in the coding region of an ORF, substitution of one codon for another which encodes the same amino acid is expressly contemplated.

Any specific sequence disclosed herein can be readily screened for errors by resequencing a particular fragment, such as an ORF, in both directions (i.e., sequence both strands). Alternatively, error screening can be performed by sequencing corresponding polynucleotides of *Haemophilus influenzae* origin isolated by using part r all of the fragments in question as a probe or primer.

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO 9633276A1_I_>

Each of the ORFs of the *Haemophilus influenzae* Rd genome disclosed in Tables 1(a), 1(b) and 2, and the EMF found 5′ to the ORF, can be used in numerous ways as polynucleotide reagents. The sequences can be used as diagnostic probes or diagnostic amplification primers to detect the presence of a specific microbe, such as *Haemophilus influenzae* RD, in a sample. This is especially the case with the fragments or ORFs of Table 2, which will be highly selective for *Haemophilus influenzae*.

In addition, the fragments of the present invention, as broadly described, can be used to control gene expression through triple helix formation or antisense DNA or RNA, both of which methods are based on the binding of a polynucleotide sequence to DNA or RNA. Polynucleotides suitable for use in these methods are usually 20 to 40 bases in length and are designed to be complementary to a region of the gene involved in transcription (triple helix - see Lee *et al.*, *Nucl. Acids Res. 6*:3073 (1979); Cooney *et al.*, *Science 241*:456 (1988); and Dervan *et al.*, *Science 251*:1360 (1991)) or to the mRNA itself (antisense - Okano, *J. Neurochem. 56*:560 (1991); *Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression*, CRC Press, Boca Raton, FL (1988)).

BNSDOCID: <WO___9633276A1_I_>

Triple helix- formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA hybridization blocks translation of an mRNA molecule into polypeptide. Both techniques have been demonstrated to be effective in model systems. Information contained in the sequences of the present invention is necessary for the design of an antisense or triple helix oligonucleotide.

The present invention further provides recombinant constructs comprising one or more fragments of the *Haemophilus influenzae* Rd genome of the present invention. The recombinant constructs of the present invention comprise a vector, such as a plasmid or viral vector, into which a fragment of the *Haemophilus influenzae* Rd has been inserted, in a forward or reverse orientation. In the case of a vector comprising one of the ORFs of the present invention, the vector may further comprise regulatory sequences, including for example, a promoter, operably linked to the ORF. For vectors comprising the EMFs and UMFs of the present invention, the vector may further comprise a marker sequence or heterologous ORF operably linked to the EMF or UMF. Large numbers of suitable vectors and promoters are known to those of skill in the art and are commercially available for generating the recombinant constructs of the present invention. The following vectors are provided by way of example. Bacterial: pBs, phagescript, PsiX174, pBluescript SK, pBs KS, pNH8a, pNH16a, pNH18a, pNH46a (Stratagene); pTrc99A, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia). Eukaryotic: pWLneo, pSV2cat, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, pSVL (Pharmacia).

Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers. Two appropriate vectors are pKK232-8 and pCM7. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda P_R, and trc. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and prom ter is well within the level of ordinary skill in th art.

The present invention further provides h st cells containing any one of the isolated fragments of the *Haemophilus influenzae* Rd genome of the present invention, wherein the fragment has been introduced into the host cell using known transformulation methods. The host cell can be a higher eukaryotic host cell, such as a mammalian cell, a lower eukaryotic host cell, such as a yeast cell, or the host cell can be a procaryotic cell, such as a bacterial cell. Introduction of the recombinant construct into the host cell can be effected by calcium phosphate transfection, DEAE, dextran mediated transfection, or electroporation (Davis, L. *et al.*, *Basic Methods in Molecular Biology* (1986)).

The host cells containing one of the fragments of the *Haemophilus influenzae* Rd genome of the present invention, can be used in conventional manners to produce the gene product encoded by the isolated fragment (in the case of an ORF) or can be used to produce a heterologous protein under the control of the EMF.

The present invention further provides isolated polypeptides encoded by the nucleic acid fragments of the present invention or by degenerate variants of the nucleic acid fragments of the present invention. By "degenerate variant" is intended nucleotide fragments which differ from a nucleic acid fragment of the present invention (e.g., an ORF) by nucleotide sequence but, due to the degeneracy of the Genetic Code, encode an identical polypeptide sequence. Preferred nucleic acid fragments of the present invention are the ORFs depicted in Table 1(a) which encode proteins.

A variety of methodologies known in the art can be utilized to obtain any one of the isolated polypeptides or proteins of the present invention. At the simplest level, the amino acid sequence can be synthesized using commercially available peptide synthesizers. This is particularly useful in producing small peptides and fragments of larger polypeptides. Fragments are useful, for example, in generating antibodies against the native polypeptide. In an alternative method, the polypeptide or protein is purified from bacterial cells which naturally produce the polypeptide or protein. One skilled in the art can readily follow kn wn methods for isolating polpeptides and proteins in

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

order to obtain ne of the isolated polypeptides or proteins of the present invention. These include, but are not limited to, immunochromatography, HPLC, size-exclusion chromatography, ion-exchange chromatography, and immuno-affinity chromatography.

The polypeptides and proteins of the present invention can alternatively be purified from cells which have been altered to express the desired polypeptide or protein. As used herein, a cell is said to be altered to express a desired polypeptide or protein when the cell, through genetic manipulation, is made to produce a polypeptide or protein which it normally does not produce or which the cell normally produces at a lower level. One skilled in the art can readily adapt procedures for introducing and expressing either recombinant or synthetic sequences into eukaryotic or prokaryotic cells in order to generate a cell which produces one of the polypeptides or proteins of the present invention.

Any host/vector system can be used to express one or more of the ORFs of the present invention. These include, but are not limited to, eukaryotic hosts such as HeLa cells, Cv-1 cell, COS cells, and Sf9 cells, as well as prokaryotic host such as *E. coli* and *B. subtilis*. The most preferred cells are those which do not normally express the particular polypeptide or protein or which expresses the polypeptide or protein at low natural level.

"Recombinant," as used herein, means that a polypeptide or protein is derived from recombinant (e.g., microbial or mammalian) expression systems. "Microbial" refers to recombinant polypeptides or proteins made in bacterial or fungal (e.g., yeast) expression systems. As a product, "recombinant microbial" defines a polypeptide or protein essentially free of native endogenous substances and unaccompanied by associated native glycosylation. Polypeptides or proteins expressed in most bacterial cultures, e.g., *E. coli*, will be free of glycosylation modifications; polypeptides or proteins expressed in yeast will have a glycosylation pattern different from that expressed in mammalian cells.

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO 9633276A1 I >

"Nucleotide sequence" refers to a heteropolymer of deoxyribonucleotides. Generally, DNA segments encoding the polypeptides and proteins provided by this invention are assembled from fragments of the *Haemophilus influenzae* Rd genome and short oligonucleotide linkers, or from a series of oligonucleotides, to provide a synthetic gene which is capable of being expressed in a recombinant transcriptional unit comprising regulatory elements derived from a microbial or viral operon.

"Recombinant expression vehicle or vector" refers to a plasmid or phage or virus or vector, for expressing a polypeptide from a DNA (RNA) sequence. The expression vehicle can comprise a transcriptional unit comprising an assembly of (1) a genetic element or elements having a regulatory role in gene expression, for example, promoters or enhancers, (2) a structural or coding sequence which is transcribed into mRNA and translated into protein, and (3) appropriate transcription initiation and termination sequences. Structural units intended for use in yeast or eukaryotic expression systems preferably include a leader sequence enabling extracellular secretion of translated protein by a host cell. Alternatively, where recombinant protein is expressed without a leader or transport sequence, it may include an N-terminal methionine residue. This residue may or may not be subsequently cleaved from the expressed recombinant protein to provide a final product.

"Recombinant expression system" means host cells which have stably integrated a recombinant transcriptional unit into chromosomal DNA or carry the recombinant transcriptional unit extra chromosomally. The cells can be prokaryotic or eukaryotic. Recombinant expression systems as defined herein will express heterologous polypeptides or proteins upon induction of the regulatory elements linked to the DNA segment or synthetic gene to be expressed.

Mature proteins can be expressed in mammalian cells, yeast, bacteria, or other cells under the control of appropriate promoters. Cell-free translation systems can also be employed to produce such proteins using RNAs derived from the DNA constructs of the present invention. Appropriate cloning and

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

expression vectors for use with prokaryatic and eukaryotic hosts are described by Sambrook, *et al.*, in *Molecular Cloning: A Laboratory Manual*, Second Edition, Cold Spring Harbor, New York (1989), the disclosure of which is hereby incorporated by reference.

Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, e.g., the ampicillin resistance gene of *E. coli* and *S. cerevisiae* TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), a-factor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably, a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium. Optionally, the heterologous sequence can encode a fusion protein including an N-terminal identification peptide imparting desired characteristics, e.g., stabilization or simplified purification of expressed recombinant product.

Useful expression vectors for bacterial use are constructed by inserting a structural DNA sequence encoding a desired protein together with suitable translation initiation and termination signals in operable reading phase with a functional promoter. The vector will comprise one or more phenotypic selectable markers and an origin of replication to ensure maintenance of the vector and to, if desirable, provide amplification within the host. Suitable prokaryotic hosts for transformation include *E. coli*, *Bacillus subtilis*, *Salmonella typhimurium* and various species within the genera Pseudomonas, Streptomyces, and Staphylococcus, although others may, also be employed as a matter of choice.

As a representative but nonlimiting example, useful expression vectors for bacterial use can comprise a selectable marker and bacterial origin of replication derived from commercially available plasmids comprising genetic

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

elements of the well kn wn cloning vector pBR322 (ATCC 37017). Such commercial vectors include, for example, pKK223-3 (Pharmacia Fine Chemicals, Uppsala, Sweden) and GEM 1 (Promega Biotec, Madison, WI, USA). These pBR322 "backbone" sections are combined with an appropriate promoter and the structural sequence to be expressed.

Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter is derepressed by appropriate means (e.g., temperature shift or chemical induction) and cells are cultured for an additional period. Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described by Gluzman, *Cell 23*:175 (1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5′ flanking nontranscribed sequences. DNA sequences derived from the SV40 viral genome, for example, SV40 origin, early promoter, enhancer, splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

Recombinant polypeptides and proteins produced in bacterial culture is usually isolated by initial extraction from cell pellets, followed by one or more salting-out, aqueous ion exchange or size exclusion chromatography steps. Protein refolding steps can be used, as necessary, in completing configuration of the mature protein. Finally, high performance liquid chromatography (HPLC) can be employed for final purification steps. Microbial cells employed in expression of proteins can be disrupted by any

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents.

The present invention further includes isolated polypeptides, proteins and nucleic acid molecules which are substantially equivalent to those herein described. As used herein, substantially equivalent can refer both to nucleic acid and amino acid sequences, for example a mutant sequence, that varies from a reference sequence by one or more substitutions, deletions, or additions, the net effect of which does not result in an adverse functional dissimilarity between reference and subject sequences. For purposes of the present invention, sequences having equivalent biological activity, and equivalent expression characteristics are considered substantially equivalent. For purposes of determining equivalence, truncation of the mature sequence should be disregarded.

The invention further provides methods of obtaining homologs from other strains of *Haemophilus influenzae*, of the fragments of the *Haemophilus influenzae* Rd genome of the present invention and homologs of the proteins encoded by the ORFs of the present invention. As used herein, a sequence or protein of *Haemophilus influenzae* is defined as a homolog of a fragment of the *Haemophilus influenzae* Rd genome or a protein encoded by one of the ORFs of the present invention, if it shares significant homology to one of the fragments of the *Haemophilus influenzae* Rd genome of the present invention or a protein encoded by one of the ORFs of the present invention. Specifically, by using the sequence disclosed herein as a probe or as primers, and techniques such as PCR cloning and colony/plaque hybridization, one skilled in the art can obtain homologs.

As used herein, two nucleic acid molecules or proteins are said to "share significant homology" if the two contain regions which process greater than 85% sequence (amino acid or nucleic acid) homology.

Region specific primers or probes derived from the nucleotide sequence provided in SEQ ID NO:1 or from a nucleotide sequence at least 99.9% identical to SEQ ID NO:1 can be used to prime DNA synthesis and PCR amplification, as well as to identify colonies containing cloned DNA encoding

SUBSTITUTE SHEET (RULE 26)

3NSDOCID: <WO___9633276A1_I_>

a homolog using known methods (Innis *et al.*, *PCR Protocols*, Academic Press, San Diego, CA (1990)).

When using primers derived from SEQ ID NO:1 or from a nucleotide sequence at least 99.9% identical to SEQ ID NO:1, one skilled in the art will recognize that by employing high stringency conditions (e.g., annealing at 50-60°C) only sequences which are greater than 75% homologous to the primer will be amplified. By employing lower stringency conditions (e.g., annealing at 35-37°C), sequences which are greater than 40-50% homologous to the primer will also be amplified.

When using DNA probes derived from SEQ ID NO:1 or from a nucleotide sequence at least 99.9% identical to SEQ ID NO:1 for colony/plaque hybridization, one skilled in the art will recognize that by employing high stringency conditions (e.g., hybridizing at 50-65°C in 5X SSPC and 50% formamide, and washing at 50-65°C in 0.5X SSPC), sequences having regions which are greater than 90% homologous to the probe can be obtained, and that by employing lower stringency conditions (e.g., hybridizing at 35-37°C in 5X SSPC and 40-45% formamide, and washing at 42°C in SSPC), sequences having regions which are greater than 35-45% homologous to the probe will be obtained.

Any organism can be used as the source for homologs of the present invention so long as the organism naturally expresses such a protein or contains genes encoding the same. The most preferred organism for isolating homologs are bacterias which are closely related to *Haemophilus influenzae* Rd.

Uses for the Compositions of the Invention

Each ORF provided in Table 1(a) was assigned to one of 102 biological role categories adapted from Riley, M., *Microbiology Reviews 57(4)*:862 (1993)). This allows the skilled artisan to determine a use for each identified coding sequence. Tables 1(a) further provides an identification of the type of polypeptide which is encoded for by each ORF. As a result, one skilled in the art can use the polypeptides of the present invention for commercial,

therapeutic and industrial purposes consistent with the type of putative identificati n of the polypeptide.

Such identifications permit one skilled in the art to use the *Haemophilus influenzae* ORFs in a manner similar to the known type of sequences for which the identification is made; for example, to ferment a particular sugar source or to produce a particular metabolite. (For a review of enzymes used within the commercial industry, *see Biochemical Engineering and Biotechnology Handbook* 2nd, eds. Macmillan Publ. Ltd., NY (1991) and Biocatalysts in Organic Syntheses, ed. J. Tramper *et al.*, Elsevier Science Publishers, Amsterdam, The Netherlands (1985)).

1. Biosynthetic Enzymes

Open reading frames encoding proteins involved in mediating the catalytic reactions involved in intermediary and macromolecular metabolism, the biosynthesis of small molecules, cellular processes and other functions includes enzymes involved in the degradation of the intermediary products of metabolism, enzymes involved in central intermediary metabolism, enzymes involved in respiration, both aerobic and anaerobic, enzymes involved in fermentation, enzymes involved in ATP proton motor force conversion, enzymes involved in broad regulatory function, enzymes involved in amino acid synthesis, enzymes involved in nucleotide synthesis, enzymes involved in cofactor and vitamin synthesis, can be used for industrial biosynthesis. The various metabolic pathways present in *Haemophilus* can be identified based on absolute nutritional requirements as well as by examining the various enzymes identified in Table 1(a).

Identified within the category of intermediary metabolism, a number of the proteins encoded by the identified ORFs in Tables 1(a) are particularly involved in the degradation of intermediary metabolites as well as non-macromolecular metabolism. Some of the enzymes identified include amylases, glucose oxidases, and catalase.

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO__9633276A1_I_>

Proteolytic enzymes are another class of commercially important enzymes. Proteolytic enzymes find use in a number of industrial processes including the processing of flax and other vegetable fibers, in the extraction, clarification and depectinization of fruit juices, in the extraction of vegetables' oil and in the maceration of fruits and vegetables to give unicellular fruits. A detailed review of the proteolytic enzymes used in the food industry is provided by Rombouts *et al.*, *Symbiosis 21*:79 (1986) and Voragen *et al.* in *Biocatalyst in Agricultural Biotechnology*, edited J.R. Whitaker *et al.*, *American Chemical Society Symposium Series 389*:93 (1989)).

The metabolism of glucose, galactose, fructose and xylose are important parts of the primary metabolism of *Haemophilus*. Enzymes involved in the degradation of these sugars can be used in industrial fermentation. Some of the important sugar transforming enzymes, from a commercial viewpoint, include sugar isomerases such as glucose isomerase. Other metabolic enzymes have found commercial use such as glucose oxidases which produces ketogulonic acid (KGA). KGA is an intermediate in the commercial production of ascorbic acid using the Reichstein's procedure (*see* Krueger *et al.*, *Biotechnology 6(A)*, Rhine, H.J. *et al.*, eds., Verlag Press, Weinheim, Germany (1984)).

Glucose oxidase (GOD) is commercially available and has been used in purified form as well as in an immobilized form for the deoxygenation of beer. *See* Hartmeir *et al.*, *Biotechnology Letters 1*:21 (1979). The most important application of GOD is the industrial scale fermentation of gluconic acid. Market for gluconic acids which are used in the detergent, textile, leather, photographic, pharmaceutical, food, feed and concrete industry (*see* Bigelis in *Gene Manipulations and Fungi*, Benett, J.W. *et al.*, eds., Academic Press, New York (1985), p. 357). In addition to industrial applications, GOD has found applications in medicine for quantitative determination of glucose in body fluids recently in biotechnology for analyzing syrups from starch and cellulose hydrosylates. *See* Owusu *et al.*, *Biochem. et Biophysica. Acta. 872*:83 (1986).

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

The main sweetener used in the world today is sugar which comes from sugar beets and sugar cane. In the field of industrial enzymes, the glucose isomerase process shows the largest expansion in the market today. Initially, soluble enzymes were used and later immobilized enzymes were developed (Krueger *et al.*, *Biotechnology, The Textbook of Industrial Microbiology*, Sinauer Associated Incorporated, Sunderland, Massachusetts (1990)). Today, the use of glucose-produced high fructose syrups is by far the largest industrial business using immobilized enzymes. A review of the industrial use of these enzymes is provided by Jorgensen, *Starch 40*:307 (1988).

Proteinases, such as alkaline serine proteinases, are used as detergent additives and thus represent one of the largest volumes of microbial enzymes used in the industrial sector. Because of their industrial importance, there is a large body of published and unpublished information regarding the use of these enzymes in industrial processes. (See Faultman *et al.*, *Acid Proteases Structure Function and Biology*, Tang, J., ed., Plenum Press, New York (1977) and Godfrey *et al.*, *Industrial Enzymes*, MacMillan Publishers, Surrey, UK (1983) and Hepner *et al.*, *Report Industrial Enzymes by 1990*, Hel Hepner & Associates, London (1986)).

Another class of commercially usable proteins of the present invention are the microbial lipases identified in Table 1 (see Macrae *et al.*, *Philosophical Transactions of the Chiral Society of London 310*:227 (1985) and Poserke, *Journal of the American Oil Chemist Society 61*:1758 (1984). A major use of lipases is in the fat and oil industry for the production of neutral glycerides using lipase catalyzed inter-esterification of readily available triglycerides. Application of lipases include the use as a detergent additive to facilitate the removal of fats from fabrics in the course of the washing procedures.

The use of enzymes, and in particular microbial enzymes, as catalyst for key steps in the synthesis of complex organic molecules is gaining popularity at a great rate. One area of great interest is the preparation of chiral intermediates. Preparation of chiral intermediates is of interest to a wide range of synthetic chemists particularly those scientists involved with the preparation

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

of new pharmaceuticals, agrochemicals, fragrances and flavors. (See Davies *et al.*, *Recent Advances in the Generation of Chiral Intermediates Using Enzymes*, CRC Press, Boca Raton, Florida (1990)). The following reactions catalyzed by enzymes are of interest to organic chemists: hydrolysis of carboxylic acid esters, phosphate esters, amides and nitriles, esterification reactions, trans-esterification reactions, synthesis of amides, reduction of alkanones and oxoalkanates, oxidation of alcohols to carbonyl compounds, oxidation of sulfides to sulfoxides, and carbon bond forming reactions such as the aldol reaction. When considering the use of an enzyme encoded by one of the ORFs of the present invention for biotransformation and organic synthesis it is sometimes necessary to consider the respective advantages and disadvantages of using a microorganism as opposed to an isolated enzyme. Pros and cons of using a whole cell system on the one hand or an isolated partially purified enzyme on the other hand, has been described in detail by Bud *et al.*, *Chemistry in Britain* (1987), p. 127.

Amino transferases, enzymes involved in the biosynthesis and metabolism of amino acids, are useful in the catalytic production of amino acids. The advantages of using microbial based enzyme systems is that the amino transferase enzymes catalyze the stereo-selective synthesis of only *l*-amino acids and generally possess uniformly high catalytic rates. A description of the use of amino transferases for amino acid production is provided by Roselle-David, *Methods of Enzymology 136*:479 (1987).

Another category of useful proteins encoded by the ORFs of the present invention include enzymes involved in nucleic acid synthesis, repair, and recombination. A variety of commercially important enzymes have previously been isolated from members of *Haemophilus* sp. These include the Hinc II, Hind III, and Hinf I restriction endonucleases. Table 1(a) identifies a wide array of enzymes, such as restriction enzymes, ligases, gyrases and methylases, which have immediate use in the biotechnology industry.

2. Generation of Antibodies

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

As described here, the proteins of the present invention, as well as h mologs thereof, can be used in a variety procedures and methods known in the art which are currently applied to other proteins. The proteins of the present invention can further be used to generate an antibody which selectively binds the protein. Such antibodies can be either monoclonal or polyclonal antibodies, as well fragments of these antibodies, and humanized forms.

The invention further provides antibodies which selectively bind to one of the proteins of the present invention and hybridomas which produce these antibodies. A hybridoma is an immortalized cell line which is capable of secreting a specific monoclonal antibody.

In general, techniques for preparing polyclonal and monoclonal antibodies as well as hybridomas capable of producing the desired antibody are well known in the art (Campbell, A.M., *Monoclonal Antibody Technology: Laboratory Techniques in Biochemistry and Molecular Biology*, Elsevier Science Publishers, Amsterdam, The Netherlands (1984); St. Groth *et al.*, *J. Immunol. Methods* 35:1-21 (1980); Kohler and Milstein, *Nature* 256:495-497 (1975)), the trioma technique, the human B-cell hybridoma technique (Kozbor *et al.*, *Immunology Today* 4:72 (1983); Cole *et al.*, in *Monoclonal Antibodies and Cancer Therapy*, Alan R. Liss, Inc. (1985), pp. 77-96).

Any animal (mouse, rabbit, etc.) which is known to produce antibodies can be immunized with the pseudogene polypeptide. Methods for immunization are well known in the art. Such methods include subcutaneous or interperitoneal injection of the polypeptide. One skilled in the art will recognize that the amount of the protein encoded by the ORF of the present invention used for immunization will vary based on the animal which is immunized, the antigenicity of the peptide and the site of injection.

The protein which is used as an immunogen may be modified or administered in an adjuvant in order to increase the protein's antigenicity. Methods of increasing the antigenicity of a protein are well known in the art and include, but are not limited to coupling the antigen with a heterologous

protein (such as globulin or β-galactosidase) or through the inclusion of an adjuvant during immunization.

For monoclonal antibodies, spleen cells from the immunized animals are removed, fused with myeloma cells, such as SP2/0-Ag14 myeloma cells, and allowed to become monoclonal antibody producing hybridoma cells.

Any one of a number of methods well known in the art can be used to identify the hybridoma cell which produces an antibody with the desired characteristics. These include screening the hybridomas with an ELISA assay, western blot analysis, or radioimmunoassay (Lutz *et al.*, *Exp. Cell Res.* *175*:109-124 (1988)).

Hybridomas secreting the desired antibodies are cloned and the class and subclass is determined using procedures known in the art (Campbell, A.M., *Monoclonal Antibody Technology: Laboratory Techniques in Biochemistry and Molecular Biology*, Elsevier Science Publishers, Amsterdam, The Netherlands (1984)).

Techniques described for the production of single chain antibodies (U.S. Patent 4,946,778) can be adapted to produce single chain antibodies to proteins of the present invention.

For polyclonal antibodies, antibody containing antisera is isolated from the immunized animal and is screened for the presence of antibodies with the desired specificity using one of the above-described procedures.

The present invention further provides the above-described antibodies in detectably labelled form. Antibodies can be detectably labelled through the use of radioisotopes, affinity labels (such as biotin, avidin, etc.), enzymatic labels (such as horseradish peroxidase, alkaline phosphatase, etc.) fluorescent labels (such as FITC or rhodamine, etc.), paramagnetic atoms, etc. Procedures for accomplishing such labelling are well-known in the art, for example see (Sternberger, L.A. *et al.*, *J. Histochem. Cytochem.* *18*:315 (1970); Bayer, E.A. *et al.*, *Meth. Enzym.* *62*:308 (1979); Engval, E. *et al.*, *Immunol.* *109*:129 (1972); Goding, J.W. *J. Immunol. Meth.* *13*:215 (1976)).

BNSDOCID: <WO___9633276A1_I_>

The labeled antibodies of the present invention can be used for *in vitro*, *in vivo*, and *in situ* assays to identify cells or tissues in which a fragment of the *Haemophilus influenzae* Rd genome is expressed.

The present invention further provides the above-described antibodies immobilized on a solid support. Examples of such solid supports include plastics such as polycarbonate, complex carbohydrates such as agarose and sepharose, acrylic resins and such as polyacrylamide and latex beads. Techniques for coupling antibodies to such solid supports are well known in the art (Weir, D.M. *et al.*, "*Handbook of Experimental Immunology*" 4th Ed., Blackwell Scientific Publications, Oxford, England, Chapter 10 (1986); Jacoby, W.D. *et al.*, *Meth. Enzym. 34* Academic Press, N.Y. (1974)). The immobilized antibodies of the present invention can be used for *in vitro*, *in vivo*, and *in situ* assays as well as for immunoaffinity purification of the proteins of the present invention.

3. Diagnostic Assays and Kits

The present invention further provides methods to identify the expression of one of the ORFs of the present invention, or homolog thereof, in a test sample, using one of the DFs or antibodies of the present invention.

In detail, such methods comprise incubating a test sample with one or more of the antibodies or one or more of the DFs of the present invention and assaying for binding of the DFs or antibodies to components within the test sample.

Conditions for incubating a DF or antibody with a test sample vary. Incubation conditions depend on the format employed in the assay, the detection methods employed, and the type and nature of the DF or antibody used in the assay. One skilled in the art will recognize that any one of the commonly available hybridization, amplification or immunological assay formats can readily be adapted to employ the DFs r antibodies of the present invention. Examples of such assays can be found in Chard, T., *An*

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

Introduction to Radioimmunoassay and Related Techniques, Elsevier Science Publishers, Amsterdam, The Netherlands (1986); Bullock, G.R. *et al.*, *Techniques in Immunocytochemistry*, Academic Press, Orlando, FL Vol. 1 (1982), Vol. 2 (1983), Vol. 3 (1985); Tijssen, P., *Practice and Theory of Enzyme Immunoassays: Laboratory Techniques in Biochemistry and Molecular Biology*, Elsevier Science Publishers, Amsterdam, The Netherlands (1985).

The test samples of the present invention include cells, protein or membrane extracts of cells, or biological fluids such as sputum, blood, serum, plasma, or urine. The test sample used in the above-described method will vary based on the assay format, nature of the detection method and the tissues, cells or extracts used as the sample to be assayed. Methods for preparing protein extracts or membrane extracts of cells are well known in the art and can be readily be adapted in order to obtain a sample which is compatible with the system utilized.

In another embodiment of the present invention, kits are provided which contain the necessary reagents to carry out the assays of the present invention.

Specifically, the invention provides a compartmentalized kit to receive, in close confinement, one or more containers which comprises: (a) a first container comprising one of the DFs or antibodies of the present invention; and (b) one or more other containers comprising one or more of the following: wash reagents, reagents capable of detecting presence of a bound DF or antibody.

In detail, a compartmentalized kit includes any kit in which reagents are contained in separate containers. Such containers include small glass containers, plastic containers or strips of plastic or paper. Such containers allows one to efficiently transfer reagents from one compartment to another compartment such that the samples and reagents are not cross-contaminated, and the agents or solutions of each container can be added in a quantitative fashion from one compartment to another. Such containers will include a container which will accept the test sample, a container which contains the

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

antibodies used in the assay, containers which contain wash reagents (such as phosphate buffered saline, Tris-buffers, etc.), and containers which contain the reagents used to detect the bound antibody or DF.

Types of detection reagents include labelled nucleic acid probes, labelled secondary antibodies, or in the alternative, if the primary antibody is labelled, the enzymatic, or antibody binding reagents which are capable of reacting with the labelled antibody. One skilled in the art will readily recognize that the disclosed DFs and antibodies of the present invention can be readily incorporated into one of the established kit formats which are well known in the art.

4. Screening Assay for Binding Agents

Using the isolated proteins of the present invention, the present invention further provides methods of obtaining and identifying agents which bind to a protein encoded by one of the ORFs of the present invention or to one of the fragments and the *Haemophilus* genome herein described.

In detail, said method comprises the steps of:

(a) contacting an agent with an isolated protein encoded by one of the ORFs of the present invention, or an isolated fragment of the *Haemophilus* genome; and

(b) determining whether the agent binds to said protein or said fragment.

The agents screened in the above assay can be, but are not limited to, peptides, carbohydrates, vitamin derivatives, or other pharmaceutical agents. The agents can be selected and screened at random or rationally selected or designed using protein modeling techniques.

For random screening, agents such as peptides, carbohydrates, pharmaceutical agents and the like are selected at random and are assayed for their ability t bind t the protein encoded by the ORF of the present invention.

Alternatively, agents may be rationally selected or designed. As used herein, an agent is said to be "rationally selected or designed" when the agent is chosen based on the configuration of the particular protein. For example, one skilled in the art can readily adapt currently available procedures to generate peptides, pharmaceutical agents and the like capable of binding to a specific peptide sequence in order to generate rationally designed antipeptide peptides, for example see Hurby *et al.*, Application of Synthetic Peptides: Antisense Peptides," In *Synthetic Peptides, A User's Guide*, W.H. Freeman, NY (1992), pp. 289-307, and Kaspczak *et al.*, *Biochemistry 28*:9230-8 (1989), or pharmaceutical agents, or the like.

In addition to the foregoing, one class of agents of the present invention, as broadly described, can be used to control gene expression through binding to one of the ORFs or EMFs of the present invention. As described above, such agents can be randomly screened or rationally designed/selected. Targeting the ORF or EMF allows a skilled artisan to design sequence specific or element specific agents, modulating the expression of either a single ORF or multiple ORFs which rely on the same EMF for expression control.

One class of DNA binding agents are agents which contain base residues which hybridize or form a triple helix formation by binding to DNA or RNA. Such agents can be based on the classic phosphodiester, ribonucleic acid backbone, or can be a variety of sulfhydryl or polymeric derivatives which have base attachment capacity.

Agents suitable for use in these methods usually contain 20 to 40 bases and are designed to be complementary to a region of the gene involved in transcription (triple helix - see Lee *et al.*, *Nucl. Acids Res. 6*:3073 (1979); Cooney *et al.*, *Science 241*:456 (1988); and Dervan *et al.*, *Science 251*: 1360 (1991)) or to the mRNA itself (antisense - Okano, *J. Neurochem. 56*:560 (1991); *Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression*, CRC Press, Boca Raton, FL (1988)). Triple helix- formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA

hybridization blocks translation of an mRNA molecule into polypeptide. Both techniques have been demonstrated to be effective in model systems. Information contained in the sequences of the present invention is necessary for the design of an antisense or triple helix oligonucleotide and other DNA binding agents.

Agents which bind to a protein encoded by one of the ORFs of the present invention can be used as a diagnostic agent, in the control of bacterial infection by modulating the activity of the protein encoded by the ORF. Agents which bind to a protein encoded by one of the ORFs of the present invention can be formulated using known techniques to generate a pharmaceutical composition for use in controlling *Haemophilus* growth and infection.

5. Vaccine and Pharmaceutical Composition

The present invention further provides pharmaceutical agents which can be used to modulate the growth of *Haemophilus influenzae*, or another related organism, *in vivo* or *in vitro*. As used herein, a "pharmaceutical agent" is defined as a composition of matter which can be formulated using known techniques to provide a pharmaceutical compositions. As used herein, the "pharmaceutical agents of the present invention" refers the pharmaceutical agents which are derived from the proteins encoded by the ORFs of the present invention or are agents which are identified using the herein described assays.

As used herein, a pharmaceutical agent is said to "modulated the growth of *Haemophilus sp.*, or a related organism, *in vivo* or *in vitro*," when the agent reduces the rate of growth, rate of division, or viability of the organism in question. The pharmaceutical agents of the present invention can modulate the growth of an organism in many fashions, although an understanding of the underlying mechanism of action is not needed to practice the use of the pharmaceutical agents of the present invention. Some agents will modulate the growth by binding to an important protein thus blocking the

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO__9633276A1_I_>

biological activity of the protein, while other agents may bind to a component f the outer surface f the organism blocking attachment or rendering the organism more prone to act the bodies nature immune system. Alternatively, the agent may be comprise a protein encoded by one of the ORFs of the present invention and serve as a vaccine. The development and use of a vaccine based on outer membrane components, such as the LPS, are well known in the art.

As used herein, a "related organism" is a broad term which refers to any organism whose growth can be modulated by one of the pharmaceutical agents of the present invention. In general, such an organism will contain a homolog of the protein which is the target of the pharmaceutical agent or the protein used as a vaccine. As such, related organism do not need to be bacterial but may be fungal or viral pathogens.

The pharmaceutical agents and compositions of the present invention may be administered in a convenient manner such as by the oral, topical, intravenous, intraperitoneal, intramuscular, subcutaneous, intranasal or intradermal routes. The pharmaceutical compositions are administered in an amount which is effective for treating and/or prophylaxis of the specific indication. In general, they are administered in an amount of at least about 10 μg/kg body weight and in most cases they will be administered in an amount not in excess of about 8 mg/Kg body weight per day. In most cases, the dosage is from about 10 μg/kg to about 1 mg/kg body weight daily, taking into account the routes of administration, symptoms, etc.

The agents of the present invention can be used in native form or can be modified to form a chemical derivative. As used herein, a molecule is said to be a "chemical derivative" of another molecule when it contains additional chemical moieties not normally a part of the molecule. Such moieties may improve the molecule's solubility, absorption, biological half life, etc. The moieties may alternatively decrease the toxicity of the molecule, eliminate or attenuate any undesirable side effect of the molecule, etc. Moieties capable of

mediating such effects are disclosed in *Remington's Pharmaceutical Sciences* (1980).

For example, a change in the immunological character of the functional derivative, such as affinity for a given antibody, is measured by a competitive type immunoassay. Changes in immunomodulation activity are measured by the appropriate assay. Modifications of such protein properties as redox or thermal stability, biological half-life, hydrophobicity, susceptibility to proteolytic degradation or the tendency to aggregate with carriers or into multimers are assayed by methods well known to the ordinarily skilled artisan.

The therapeutic effects of the agents of the present invention may be obtained by providing the agent to a patient by any suitable means (i.e., inhalation, intravenously, intramuscularly, subcutaneously, enterally, or parenterally). It is preferred to administer the agent of the present invention so as to achieve an effective concentration within the blood or tissue in which the growth of the organism is to be controlled.

To achieve an effective blood concentration, the preferred method is to administer the agent by injection. The administration may be by continuous infusion, or by single or multiple injections.

In providing a patient with one of the agents of the present invention, the dosage of the administered agent will vary depending upon such factors as the patient's age, weight, height, sex, general medical condition, previous medical history, etc. In general, it is desirable to provide the recipient with a dosage of agent which is in the range of from about 1 pg/kg to 10 mg/kg (body weight of patient), although a lower or higher dosage may be administered. The therapeutically effective dose can be lowered by using combinations of the agents of the present invention or another agent.

As used herein, two or more compounds or agents are said to be administered "in combination" with each other when either (1) the physiological effects of each compound, or (2) the serum concentrations of each compound can be measured at the same time. The composition of the

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO__ 9633276A1 I >

present invention can be administered concurrently with, prior to, or following the administration of the other agent.

The agents of the present invention are intended to be provided to recipient subjects in an amount sufficient to decrease the rate of growth (as defined above) of the target organism.

The administration of the agent(s) of the invention may be for either a "prophylactic" or "therapeutic" purpose. When provided prophylactically, the agent(s) are provided in advance of any symptoms indicative of the organisms growth. The prophylactic administration of the agent(s) serves to prevent, attenuate, or decrease the rate of onset of any subsequent infection. When provided therapeutically, the agent(s) are provided at (or shortly after) the onset of an indication of infection. The therapeutic administration of the compound(s) serves to attenuate the pathological symptoms of the infection and to increase the rate of recovery.

The agents of the present invention are administered to the mammal in a pharmaceutically acceptable form and in a therapeutically effective concentration. A composition is said to be "pharmacologically acceptable" if its administration can be tolerated by a recipient patient. Such an agent is said to be administered in a "therapeutically effective amount" if the amount administered is physiologically significant. An agent is physiologically significant if its presence results in a detectable change in the physiology of a recipient patient.

The agents of the present invention can be formulated according to known methods to prepare pharmaceutically useful compositions, whereby these materials, or their functional derivatives, are combined in admixture with a pharmaceutically acceptable carrier vehicle. Suitable vehicles and their formulation, inclusive of other human proteins, e.g., human serum albumin, are described, for example, in *Remington's Pharmaceutical Sciences* (16th ed., Osol, A., Ed., Mack, Easton PA (1980)). In order to form a pharmaceutically acceptable composition suitable for effective administrati n, such compositions

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

will contain an effective amount of one or more of the agents of the present invention, together with a suitable amount of carrier vehicle.

Additional pharmaceutical methods may be employed to control the duration of action. Control release preparations may be achieved through the use of polymers to complex or absorb one or more of the agents of the present invention. The controlled delivery may be exercised by selecting appropriate macromolecules (for example polyesters, polyamino acids, polyvinyl, pyrrolidone, ethylenevinylacetate, methylcellulose, carboxymethylcellulose, or protamine, sulfate) and the concentration of macromolecules as well as the methods of incorporation in order to control release. Another possible method to control the duration of action by controlled release preparations is to incorporate agents of the present invention into particles of a polymeric material such as polyesters, polyamino acids, hydrogels, poly(lactic acid) or ethylene vinylacetate copolymers. Alternatively, instead of incorporating these agents into polymeric particles, it is possible to entrap these materials in microcapsules prepared, for example, by coacervation techniques or by interfacial polymerization, for example, hydroxymethylcellulose or gelatine-microcapsules and poly(methylmethacylate) microcapsules, respectively, or in colloidal drug delivery systems, for example, liposomes, albumin microspheres, microemulsions, nanoparticles, and nanocapsules or in macroemulsions. Such techniques are disclosed in *Remington's Pharmaceutical Sciences* (1980).

The invention further provides a pharmaceutical pack or kit comprising one or more containers filled with one or more of the ingredients of the pharmaceutical compositions of the invention. Associated with such container(s) can be a notice in the form prescribed by a governmental agency regulating the manufacture, use or sale of pharmaceuticals or biological products, which notice reflects approval by the agency of manufacture, use or sale for human administration. In addition, the agents of the present invention may be employed in conjunction with other therapeutic compounds.

SUBSTITUTE SHEET (RULE 26)

SNSDOCID: <WO___9633276A1_I_>

6. *Shot-Gun Approach to Megabase DNA Sequencing*

The present invention further provides the first demonstration that a sequence of greater than one megabase can be sequenced using a random shotgun approach. This procedure, described in detail in the examples that follow, has eliminated the up front cost of isolating and ordering overlapping or contiguous subclones prior to the start of the sequencing protocols.

Certain aspects of the present invention are described in greater detail in the non-limiting Examples that follow.

Examples

Experimental Design and Methods

1. Shotgun Sequencing Strategy

The overall strategy for a shotgun approach to whole genome sequencing is outlined in Table 3. The theory of shotgun sequencing follows from the Lander and Waterman (Landerman and Waterman, *Genomics 2*: 231 (1988)) application of the equation for the Poisson distribution $p_x = m^x e^{-m}/x!$, where x is the number of occurrences of an event, m is the mean number of occurrences, and p_x is the probability that any given base is not sequenced after a certain amount of random sequence has been generated. If L is the genome length, n is the number of clone insert ends sequenced, and w is the sequencing read length, then $m = nw/L$, and the probability that no clone originates at any of the w bases preceding a given base, i.e., the probability that the base is not sequenced, is $p_0 = e^{-m}$. Using the fold coverage as the unit for m, one sees that after 1.8 Mb of sequence has been randomly generated, $m = 1$, representing 1X coverage. In this case, $p_0 = e^{-1} = .37$, thus approximately 37% is unsequenced. For example, 5X coverage (approximately 9500 clones sequenced from both insert ends and an average sequence read length of 460 bp) yields $p_0 = e^{-5} = 0.0067$, or 0.67% unsequenced. The total gap length is Le^{-m}, and the average gap size is L/n. 5X coverage would leave about 128 gaps averaging about 100 bp in size. The treatment is essentially that of Lander and Waterman, *Genomics 2*:231 (1988). Table 4 illustrates the coverage for a 1.9 Mb genome with an average fragment size of 460 bp.

2. Random Library Construction

In order to approximate the random model described above during actual sequencing, a nearly ideal library of cloned genomic fragment is required. The following library construction procedure was developed to achieve this.

H. influenzae Rd KW20 DNA was prepared by phenol extraction. A mixture (3.3 ml) containing 600 μg DNA, 300 mM sodium acetate, 10 mM Tris-HCl, 1 mM Na-EDTA, 30% glycerol was sonicated (Branson Model 450 Sonicator) at the lowest energy setting for 1 min. at 0° using a 3 mm probe. The DNA was ethanol precipitated and redissolved in 500 μl TE buffer. To create blunt-ends, a 100 μl aliquot was digested for 10 min at 30° in 200 μl BAL31 buffer with 5 units BAL31 nuclease (New England BioLabs). The DNA was phenol-extracted, ethanol-precipitated, redissolved in 100 μl TE buffer, electrophoresed on a 1.0% low melting agarose gel, and the 1.6-2.0 kb size fraction was excised, phenol-extracted, and redissolved in 20 μl TE buffer. A two-step ligation procedure was used to produce a plasmid library with 97% insert of which >99% were single inserts. The first ligation mixture (50 μl) contained 2 μg of DNA fragments, 2 μg *SmaI*/BAP pUC18 DNA (Pharmacia), and 10 units T4 ligase (GIBCO/BRL), and incubation was at 14° for 4 hr. After phenol extraction and ethanol precipitation, the DNA was dissolved in 20 μl TE buffer and electrophoresed on a 1.0% low melting agarose gel. A ladder of ethidium bromide-stained linear bands, identified by size as insert (i), vector (v), v+i, v+2i, v+3i, ... was visualized by 360 nm UV light, and the v+i DNA was excised and recovered in 20 μl TE. The v+i DNA was blunt-ended by T4 polymerase treatment for 5 min. at 37° in a reaction mixture (50 μl) containing the v+i linears, 500 μM each of the 4 dNTP's, and 9 units of T4 polymerase (New England BioLabs) under recommended buffer conditions. After phenol extraction and ethanol precipitation the repaired v+i linears were dissolved in 20 μl TE. The final ligation to produce circles was carried out in a 50 μl reaction containing 5 μl of v+i linears and 5 units of T4 ligase at 14° overnight. After 10 min. at 70° the reaction mixture was stored at -20°.

BNSDOCID: <WO___9633276A1_I_>

This two-stage procedure resulted in a molecularly random collection of single-insert plasmid recombinants with minimal contamination from double-insert chimeras (<1%) or free vector (<3%). Since deviation from randomness is most likely to occur during cloning, *E. coli* host cells deficient in all recombination and restriction functions (A. Greener, *Strategies 3 (1)*:5 (1990)) were used to prevent rearrangements, deletions, and loss of clones by restriction. Transformed cells were plated directly on antibiotic diffusion plates to avoid the usual broth recovery phase which allows multiplication and selection of the most rapidly growing cells. Plating occured as follows:

A 100 μl aliquot of Epicurian Coli SURE II Supercompetent Cells (Stratagene 200152) was thawed on ice and transferred to a chilled Falcon 2059 tube on ice. A 1.7 μl aliquot of 1.42 M β-mercaptoethanol was added to the aliquot of cells to a final concentration of 25 mM. Cells were incubated on ice for 10 min. A 1 μl aliquot of the final ligation was added to the cells and incubated on ice for 30 min. The cells were heat pulsed for 30 sec. at 42° and placed back on ice for 2 min. The outgrowth period in liquid culture was eliminated from this protocol in order to minimize the preferential growth of any given transformed cell. Instead the transformation were plated directly on a nutrient rich SOB plate containing a 5 ml bottom layer of SOB agar (1.5% SOB agar: 20 g tryptone, 5 g yeast extract, 0.5 g NaCl, 1.5% Difco Agar/L). The 5 ml bottom layer is supplemented with 0.4 ml ampicillin (50 mg/ml)/100 ml SOB agar. The 15 ml top layer of SOB agar is supplemented with 1 ml X-Gal (2%), 1 ml $MgCl_2$ (1 M), and 1 ml $MgSO_4$/100 ml SOB agar. The 15 ml top layer was poured just prior to plating. Our titer was approximately 100 colonies/10 μl aliquot of transformation.

All colonies were picked for template preparation regardless of size. Only clones lost due to "poison" DNA or deleterious gene products would be deleted from the library, resulting in a slight increase in gap number over that expected.

In order to evaluate the quality of the *H. influenzae* library, sequence data were obtained from approximately 4000 templates using the M13-21 primer. The random sequence fragments were assembled using the AutoAssembler™ software (Applied Biosystems division of Perkin-Elmer (AB)) after obtaining 1300, 1800, 2500, 3200, and 3800 sequence fragments, and the number of unique assembled base pairs was determined. Based on the equations described above, an ideal plot of the number of base pairs remaining to be sequenced as a function of the # of sequenced fragments obtained with an average read length of 460 bp for a $2.5X10^6$ and a $1.9X10^6$ bp genome was determined (Figure 3). The progression of assembly was plotted using the actual data obtained from the assembly of up to 3800 sequence fragments and compared the data that is provided in the ideal plot (Figure 3). Figure 3 illustrates that there was essentially no deviation of the actual assembly data from the ideal plot, indicating that we had constructed close to an ideal random library with minimal contamination from double insert chimeras and free of vector.

3. *Random DNA Sequencing*

High quality double stranded DNA plasmid templates (19,687) were prepared using a "boiling bead" method developed in collaboration with Advanced Genetic Technology Corp. (Gaithersburg, MD) (Adams *et al.*, *Science 252*:1651 (1991); Adams *et al.*, *Nature 355*:632 (1992)). Plamid preparation was performed in a 96-well format for all stages of DNA preparation from bacterial growth through final DNA purification. Template concentration was determined using Hoechst Dye and a Millipore Cytofluor. DNA concentrations were not adjusted, but low-yielding templates were identified where possible and not sequenced. Templates were also prepared from two *H. influenzae* lambda genomic libraries. An amplified library was constructed in vector Lambda GEM-12 (Promega) and an unamplified library was constructed in Lambda DASH II (Stratagene). In particular, for the

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

unamplified lambda library, *H. influenzae* Rd KW20 DNA (> 100 kb) was partially digested in a reaction mixture (200 μl) containing 50 μg DNA, 1X *Sau*3AI buffer, 20 units Sau3AI for 6 min. at 23°. The digested DNA was phenol-extracted and electrophoresed on a 0.5% low melting agarose gel at 2V/cm for 7 hours. Fragments from 15 to 25 kb were excised and recovered in a final volume of 6 μl. One μl of fragments was used with 1 μl of DASHII vector (Stratagene) in the recommended ligation reaction. One μl of the ligation mixture was used per packaging reaction following the recommended protocol with the Gigapack II XL Packaging Extract (Stratagene, #227711). Phage were plated directly without amplification from the packaging mixture (after dilution with 500 μl of recommended SM buffer and chloroform treatment). Yield was about $2.5x10^3$ pfu/μl. The amplified library was prepared essentially as above except the lambda GEM-12 vector was used. After packaging, about $3.5x10^4$ pfu were plated on the restrictive NM539 host. The lysate was harvested in 2 ml of SM buffer and stored frozen in 7% dimethylsulfoxide. The phage titer was approximately $1x10^9$ pfu/ml.

Liquid lysates (10 ml) were prepared from randomly selected plaques and template was prepared on an anion-exchange resin (Qiagen). Sequencing reactions were carried out on plasmid templates using the AB Catalyst LabStation with Applied Biosystems PRISM Ready Reaction Dye Primer Cycle Sequencing Kits for the M13 forward (M13-21) and the M13 reverse (M13RP1) primers (Adams *et al.*, *Nature 368*:474 (1994)). Dye terminator sequencing reactions were carried out on the lambda templates on a Perkin-Elmer 9600 Thermocycler using the Applied Biosystems Ready Reaction Dye Terminator Cycle Sequencing kits. T7 and SP6 primers were used to sequence the ends of the inserts from the Lambda GEM-12 library and T7 and T3 primers were used to sequence the ends of the inserts from the Lambda DASH II library. Sequencing reactions (28,643) were performed by eight individuals using an average of fourteen AB 373 DNA Sequencers per day over a 3 month period. All sequencing reactions were analyzed using the Stretch modification of the AB 373, primarily using a 34 cm well-to-read distance. The overall

BNS page 53

sequencing success rate was 84% for M13-21 sequences, 83% for M13RP1 sequences and 65% for dye-terminator reactions. The average usable read length was 485 bp for M13-21 sequences, 444 bp for M13RP1 sequences, and 375 bp for dye-terminator reactions. Table 5 summarizes the high-throughput sequencing phase of the invention.

Richards *et al.* (Richards *et al.*, *Automated DNA sequencing and Analysis*, M.D. Adams, C. Fields, J.C. Venter, Eds. (*Academic Press, London, 1994*), Chap. 28.) described the value of using sequence from both ends of sequencing templates to facilitate ordering of contigs in shotgun assembly projects of lambda and cosmid clones. We balanced the desirability of both-end sequencing (including the reduced cost of lower total number of templates) against shorter read-lengths for sequencing reactions performed with the M13RP1 (reverse) primer compared to the M13-21 (forward) primer. Approximately one-half of the templates were sequenced from both ends. In total, 9,297 M13RP1 sequencing reactions were done. Random reverse sequencing reactions were done based on successful forward sequencing reactoins. Some M13RP1 sequences were obtained in a semi-directed fashion: M13-21 sequences pointing outward at the ends of contigs were chosen for M13RP1 sequencing in an effort to specifically order contigs. The semi-directed strategy was effective, and clone-based ordering formed an integral part of assembly and gap closure (see below).

4. Protocol for Automated Cycle Sequencing

The sequencing consisted of using eight ABI Catalyst robots and fourteen AB 373 Automated DNA Sequencers. The Catalyst robot is a publicly available sophisticated pipetting and temperature control robot which has been developed specifically for DNA sequencing reactions. The Catalyst combines pre-aliquoted templates and reaction mixes consisting of deoxy- and dideoxynucleotides, the Taq therm stable DNA polymerase, fluorescently-labelled sequencing primers, and reaction buffer. Reaction mixes and

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

templates were combined in the wells of an aluminum 96-well thermocycling plate. Thirty consecutive cycles of linear amplification (e.g., one primer synthesis) steps were performed including denaturation, annealing of primer and template, and extension of DNA synthesis. A heated lid with rubber gaskets on the thermocycling plate prevented evaporation without the need for an oil overlay.

Two sequencing protocols were used: dye-labelled primers and dye-labelled dideoxy chain terminators. The shotgun sequencing involves use of four dye-labelled sequencing primers, one for each of the four terminator nucleotide. Each dye-primer is labelled with a different fluorescent dye, permitting the four individual reactions to be combined into one lane of the 373 DNA Sequencer for electrophoresis, detection, and base-calling. AB currently supplies pre-mixed reaction mixes in bulk packages containing all the necessary non-template reagents for sequencing. Sequencing can be done with both plasmid and PCR-generated templates with both dye-primers and dye-terminators with approximately equal fidelity, although plasmid templates generally give longer usable sequences.

Thirty-two reactions were loaded per 373 Sequencer each day, for a total of 960 samples. Electrophoresis was run overnight following the manufacture's protocols, and the data was collected for twelve hours. Following electrophoresis and fluorescence detection, the AB 373 performs automatic lane tracking and base-calling. The lane-tracking was confirmed visually. Each sequence electropherogram (or fluorescence lane trace) was inspected visually and assessed for quality. Trailing sequences of low quality were removed and the sequence itself was loaded via software to a Sybase database (archived daily to a 8mm tape). Leading vector polylinker sequence was removed automatically by software program. Average edited lengths of sequences from the standard ABI 373 were around 400 bp and depended mostly on the quality of the template used for the sequencing reaction. All of the ABI 373 Sequencers were converted to Stretch Liners, which provided a

longer electrophoresis path prior to fluorescence detection, thus increasing the average number of usable bases to 500-600 bp.

Informatics

1. Data Management

A number of information management systems (LIMA) for a large-scale sequencing lab have been developed (Kerlavage *et al.*, *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*, IEEE Computer Society Press, Washington D.C., 585 (1993)). The system used to collect and assemble the sequence data was developed using the Sybase relational data management system and was designed to automate data flow whereever possible and to reduce user error. The database stores and correlates all information collected during the entire operation from template preparation to final analysis of the genome. Because the raw output of the AB 373 Sequencers was based on a Macintosh platform and the data management system chosen was based on a Unix platform, it was necessary to design and implement a variety of multi-user, client server applications which allow the raw data as well as analysis results to flow seamlessly into the database with a minimum of user effort. A description of the software programs used for large sequence assembly and managment is provided in Figure 4.

2. Assembly

An assembly engine (**TIGR Assembler**) was developed for the rapid and accurate assembly of thousands of sequence fragments. The AB **AutoAssembler™** was modified (and named **TIGR Editor**) to provide a graphical interface to the electropherogram for the purpose of editing data associated with the aligned sequence file output of **TIGR Assembler**. **TIGR Editor** maintains synchrony between the electropherogram files on the

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

Macintosh platform and the sequence data in the *H. influenzae* database on the Unix platf rm.

The **TIGR assembler** simultaneously clusters and assembles fragments of the genome. In order to obtain the speed necessary to assemble more than 10^4 fragments, the algorithm builds a hash table of 10 bp oligonucleotide subsequences to generate a list of potential sequence fragment overlaps. The number of potential overlaps for each fragment determines which fragments are likely to fall into repetitive elements. Beginning with a single seed sequence fragment, **TIGR Assembler** extends the current contig by attempting to add the best matching fragment based on oligonucleotide content. The current contig and candidate fragment are aligned using a modified version of the Smith-Waterman algorithm (Waterman, M.S., *Methods in Enzymology 164:*765 (1988)) which provides for optimal gapped alignments. The current contig is extended by the fragment only if strict criteria for the quality of the match are met. The match criteria include the minimum length of overlap, the maximum length of an unmatched end, and the minimum percentage match. These criteria are automatically lowered by the algorithm in regions of minimal coverage and raised in regions with a possible repetitive element. The number of potential overlaps for each fragment determines which fragments are likely to fall into repetitive elements. Fragments representing the boundaries of repetitive elements and potentially chimeric fragments are often rejected based on partial mismatches at the ends of alignments and excluded from the current contig. **TIGR Assembler** is designed to take advantage of clone size information coupled with sequencing from both ends of each template. It enforces the constraint that sequence fragments from two ends of the same template point toward one another in the contig and are located within a certain ranged of base pairs (definable for each clone based on the known clone size range for a given library). Assembly of 24,304 sequence fragments of *H. influenzae* required 30 hours of CPU time using one processor on a SPARCenter 2000 with 512 Mb of RAM. This process resulted in approximately 210 contigs. Because of the high stringency of the **TIGR**

SUBSTITUTE SHEET (RULE 26)

ENSDOCID <WO 9633276A1 I >

Assembler, all contigs were searched against each other using **grasta** (a modified **fasta** (Person and Lipman, *Proc. Natl. Acad. Sci. U.S.A. 85*:2444 (1988)). In this way, additional overlaps were detected which enabled compression oof the data set into 140 contigs. The location of each fragment in the contigs and extensive information about the consensus sequence itself were loaded into the *H. influenzae* relational database.

3. *Ordering Assembled Contigs*

After assembly the relative positions of the 140 contigs were unknown. The contigs were ordered by **asm_align**. **Asm_align** uses a number of relationships to identify and align contigs that are adjacent to each other. Using this algorithm, the 140 contigs were placed into 42 groups totaling 42 physical gaps (no template DNA for the region) and 98 sequence gaps (template available for gap closure).

Ordering Contigs Separated by Physical Gaps and Achieving Closure

Four integrated strategies were developed to order contigs separated by physical gaps. Oligonucleotide primers were designed and synthesized from the end of each contig group. These primers were then available for use in one or more of the strategies outlined below:

1. Southern analysis was done to develop a unique "fingerprint" for a subset of 72 of the above oligonucleotides. This procedure was based upon the supposition that labeled oligonucleotides homologous to the ends of adjacent contigs should hybridize to common DNA restriction fragments, and thus share a similar or identical hybridization pattern or "fingerprint". Oligonucleotides were labeled using 50 pmoles of each 20 mer and 250 mCi of [γ-^{32}P]ATP and T4 polynucleotide kinase. The labeled oligonucleotides were purified using Sephadex G-25 superfine (Pharmacia) and 107 cpm f each was used in a Southern hybridization analysis of *H. influenzae* Rd

SUBSTITUTE SHEET (RULE 26)

BNS pac

chromosomal DNA digested with one frequent cutters (*Ase*I) and five less frequent cutters (*Bgl*II, *Eco*RI, *Pst*I, *Xba*I, and *Pvu*II). The DNA from each digest was fractionated on a 0.7% agarose gel and transferred to Nytran Plus nylon membranes (Schleicher & Schuell). Hybridization was carried out for 16 hours at 40°. To remove non-specific signals, each blot was sequentially washed at room temperature with increasingly stringent conditions up to 0.1X SSC + 0.5% SDS. Blots were exposed to a PhosphorImager cassette (Molecular Dynamics) for several hours and hybridization patterns were visually compared.

Adjacent contigs identified in this manner were targeted for specific PCR reactions.

2. Peptide links were made by searching each contig end using **blastx** (Altschul *et al., J. Mol. Biol. 215*:403 (1990)) against a peptide database. If the ends of two contigs matched the same database sequence in an appropriate manner, then the two contigs were tentatively considered to be adjacent to each other.

3. The two lambda libraries constructed from *H. influenaze* genomic DNA were probed with oligonucleotides designed from the ends of contig groups (Kirkness *et al., Genomics 10*:985 (1991)). The positive plaques were then used to prepare templates and the sequence was determined from each end of the lambda clone insert. These sequence fragments were searched using **grasta** against a database of all contigs. Two contigs that matched the sequence from the opposite ends of the same lambda clone were ordered. The lambda clone then provided the template for closure of the sequence gap between the adjacent contigs. The lambda clones were especially valuable for solving repeat structures.

4. To confirm the order of contigs found by the other approaches and establish the order of non-ordered contigs, standard and long range (XL) PCR reactions were performed as follows.

Standard PCR was performed in the following manner. Each reaction contained a 37 μl cocktail; 16.5 μl H_2O, 3 μl 25 mM $MgCl_2$, 8 μl of a dNTP

BNSDOCID: <WO__ 9633276A1 .I_>

mix (1.25 mM each dNTP), 4.5 μl 10X PCR core buffer II (Perkin Elmer), 25 ng *H. influenzae* Rd KW20 genomic DNA. The appropriate two primers (4 μl, 3.2 pmole/μl) were added to each reaction. A hot start was performed at 95° for 5 min followed by a 75° hold. During the hold Amplitaq DNA polymerase (Perkin Elmer) 0.3 μl in 4.3 μl H_2O, 0.5 μl 10X PCR core buffer II, was added to each reaction. The PCR profile was 25 cycles of 94°/45 sec., denature; 55°/1 min., anneal; 72°/3 min, extension. All reactions were performed in a 96 well format on a Perkin Elmer GeneAmp PCR System 9600.

Long range PCR (XL PCR) was performed as follows: Each reaction contained a 35.2 μl cocktail; 12.0 μl H_2O, 2.2 μl 25 mM $Mg(OAc)_2$, 4 μl of a dNTP mix (200 μM final concentration), 12.0 μl 3.3X PCR buffer, 25 ng *H. influenzae* Rd KW20 genomic DNA. The appropriate two primers (5 μl, 3.2 pmoles/μl) was added to each reaction. A hot start was performed at 94° for 1 minute. r*Tth* polymerase, 2.0 μl (4 U/reaction) in 2.8 μl 3.3X PCR buffer II was added to each reaction. The PCR profile was 18 cycles of 94°/15 sec., denature; 62°/8 min., anneal and extend followed by 12 cycles 94°/15 sec., denature; 62°/8 min. (increase 15 sec./cycle), anneal and extend; 72°/10 min., final extension. All reactions were performed in a 96 well format on a Perkin Elmer GeneAmp PCR System 9600.

Although a PCR reaction was performed for essentially every combination of physical gap ends, techniques such as Southern fingerprinting, database matching, and the probing of large insert clones were particularly valuable in ordering contigs adjacent to each other and reducing the number of combinatorial PCR reactions necessary to achieve complete gap closure. Employing these strategies to an even greater extent in future genome projects will increase the overall efficiency of complete genome closure. The number of physical gaps ordered and closed by each of these techniques is summarized in Table 5.

Sequence information from the ends of 15-20 kb clones is particularly suitable for gap cl sure, solving repeat structures, and providing general

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

confirmation of the overall genome assembly. We were also concerned that some fragments of the *H. influenaze* gen me would be non-clonable in a high copy plasmid in *E. coli*. We reasoned that lytic lambda clones would provide the DNA for these segments. Approximately 100 random plaques were picked from the amplified lambda library, templates prepared, and sequence information obtained from each end. These sequences were searched (**grasta**) against the contigs and linked in the database to their appropriate contig, thus providing a scaffolding of lambda clones contributing additional support to the accuracy of the genome assembly (Fig 5). In addition to confirmation of the contig structure, the lambda clones provided closure for 23 physical gaps. Approximately 78% of the genome is covered by lambda clones.

Lambda clones were also useful for solving repeat structures. Repeat structures identified in the genome were small enough to be spanned by a single clone from the random insert library, except for the six ribosomal RNA operons and one repeat (2 copies) which was 5,340 bp in length. Oligonucleotide probes were designed from the unique flanks at the beginning of each repeat and hybridized to the lambda libraries. Positive plaques were identified for each flank and the sequence fragments from the ends of each clone were used to correctly orient the repeats within the genome.

The ability to distinguish and assemble the six ribosomal RNA (rRNA) operons of *H. influenaze* (16S subunit-23S subunit-5S subunit) was a test of our overall strategy to sequence and assemble a complex genome which might contain a significant number of repeat regions. The high degree of sequence similarity and the length of the six operons caused the assembly process to cluster all the underlying sequences into a few indistinguishable contigs. To determine the correct placement of the operons in the sequence, a pair of unique flanking sequences was required for each. No unique flanking sequences could be found at the left (16S rRNA) ends. This region contains the ribosomal promoter and appeared to be non-clonable in the high copy number pUC18 plasmid. However, unique sequences could be identified at the right (5S) ends. Oligonucleotide primers were designed from these six

SUBSTITUTE SHEET (RULE 26)

ENSDOCID: <WO 9633276A1 I >

flanking regions and used to probe the two lambda libraries. For each of the six rRNA operons at least one positive plaque was identified which completely spanned the rRNA operon and contained unique flanking sequence at the 16S and 5S ends. These plaques provided the templates for obtaining the unique sequence for each of the six rRNA operons.

An additional confirmation of the global structure of the assembled circular genome was obtained by comparing a computer generated restriction map based on the assembled sequence for the enzymes *Apa*l, *Sma*I, and *Rsr*II with the predicted physical map of Redfield and Lee (*Genetic Maps: locus maps of complex genomes*, S.J. O'Brien, Ed. Cold Spring Harbor Laboratory Press, New York, N.Y., 1990, 2110.). The restriction fragments from the sequence-derived map matched those from the physical map in size and relative order (Figure 5).

Editing

Simultaneous with the final gap filling process, each contig was edited visually by reassembling overlapping 10 kb sections of contigs using the AB **AutoAssembler™** and the **Fast Data Finder™** hardware. **AutoAssembler™** provides a graphical interface to electropherogram data for editing. The electropherogram data was used to assign the most likely base at each position. Where a discrepancy could not be resolved or a clear assignment made, the automatic base calls were left unchanged. Individual sequence changes were written to the electropherogram files and a replication protocol (**crash**) was used to maintain the synchrony of sequence data between the *H. influenzae* database and the electropherogram files. Following editing, contigs were reassembled with TIGR Assembler prior to annotation.

Potential frameshifts identified in the course of annotating the genome were saved as reports in the database. These reports include the coordinates in a contig which the alignment software (**praze**) predicts to be the most likely location of a missing or inserted base and a representation of the sequence

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

alignment containing the frameshift. Apparent frameshifts were used to indicate areas of the sequence which may require further editing. Frameshifts were not corrected in cases where clear electropherogram data disagreed with a frameshift. Frameshift editing was performed with **TIGR Editor**.

The rRNA and other repeat regions precluded complete assembly of the circular genome with **TIGR Assembler**. Final assembly of the genome was accomplished using **comb_asm** which splices together contigs based on short overlaps.

Accuracy of the Genome Sequence

The accuracy of the *H. influenaze* genome sequence is difficult to quantitate because there is very little previously determined *H. influenaze* sequence and most of these sequences are from other strains. There are, however, three parameters of accuracy that can be applied to the data. First, the number of apparent frameshifts in predicted *H. influenaze* genes, based on database similarities, is 148. Some of these apparent frameshifts may be in the database sequences rather than in ours, particularly considering that 49 of the apparent frameshifts are based on matches to hypothetical proteins from other organisms. Second, there are 188 bases in the genome that remain as N ambiguities (1/9,735 bp). Combining these two types of "known" errors, we can calculate a maximum sequence accuracy of 99.98%. The average coverage is 6.5X and less than 1% of the genome is single-fold coverage.

Identifying Genes

An attempt was made to predict all of the coding regions of the *H. influenzae* Rd genome and identify genes, tRNAs and rRNAs, as well as other features of the DNA sequence (e.g., repeats, regulatory sites, replication origin sites, nucleotide composition). A description of some of the readily apparent sequence features is provided below.

The *H. influenaze* Rd gen me is a circular chromosome of 1,830,121 bp. The overall G/C nucleotide content is approximately 38% (A = 31%, C = 19%, G = 19%, T = 31%, IUB = 0.035%). The G/C content of the genome was examined with several window lengths to look for global structural features. With a window of 5,000 bp, the G/C content is relatively even except for 7 large G/C-rich regions and several A/T-rich regions (Fig. 5). The G/C rich regions correspond to six rRNA operons and the location of a cryptic mu-like prophage. Genes for several proteins with similarity to proteins encoded by bacteriophage mu are located at approximately position 1.56-1.59 Mbp of the genome. This area of the genome has a markedly higher G/C content than average for *H. influenaze* (~50% G/C compared to ~38% for the rest of the genome). No significance has yet been ascertained for the source or importance of the A/T rich regions.

The minimal origin of replication (oriC) in *E. coli* is a 245 bp region defined by three copies of a thirteen base pair repeat containing a GATC core sequence at one end and four copies of a nine base pair repeat containing a TTAT core sequence at the other end. The GATC sites are methylation targets and control replication while the TTAT sites provide the binding sites for DnaA, the first step in the replication process (*Genes V*, B. Lewin Ed. (Oxford University Press, New York, 1994), chap. 18-19). An approximately 281 bp sequence (602,483 - 602,764) whose limits are defined by these same core sequences appears to define the origin of replication in *H. influenaze* Rd. These coordinates lie between sets of ribosomal operons *rrnF*, *rrnE*, *rrnD* and *rrnA*, *rrnB*, *rrnC*. These two groups of ribosomal operons are transcribed in opposite directions and the placement of the origin is consistent with their polarity for transcription. Termination of *E. coli* replication is marked by two 23 bp termination sequences located ~100 kb on either side of the midway point at which the two replication forks meet. Two potential termination sequences sharing a 10 bp core sequence with the *E. coli* termination sequence were identified in *H. influenaze* at coordinates 1,375,949-1,375,958 and 1,558,759-1,558,768. These two sets of coordinates are offset approximately

100 kb from the point 180° opposite of the proposed origin of *H. influenaze* replication.

Six rRNA operons were identified. Each rRNA operon contains three rRNA subunits and a variable spacer region in the order: 16S subunit - spacer region - 23S subunit -5S subunit. The subunit lengths are 1539 bp, 2653 bp, and 116 bp, respectively. The G/C content of the three ribosomal subunits (50%) is higher than the genome as a whole. The G/C content of the spacer region (38%) is consistent with the remainder of the genome. The nucleotide sequence of the three rRNA subunits is 100% identical in all six ribosomal operons. The rRNA operons can be grouped into two classes based on the spacer region between the 16S and 23S sequences. The shorter of the two spacer regions is 478 bp in length (*rrnB*, *rrnE*, and *rrnF*) and contains the gene for tRNA Glu. The longer spacer is 723 bp in length (*rrnA*, *rrnC*, and *rrnD*) and contains the genes for tRNA Ile and tRNA Ala. The two sets of spacer regions are also 100% identical across each group of three operons. tRNA genes are also present at the 16S and 5S ends of two of the rRNA operons. The genes for tRNA Arg, tRNA His, and tRNA Pro are located at the 16S end of *rrnE* while the genes for tRNA Trp, and tRNA Asp are located at the 5S end of *rrnA*.

The predicted coding regions of the *H. influenaze* genome were initially defined by evaluating their coding potential with the program **Genemark** (Borodovsky and McIninch, *Computers Chem. 17(2)*:123 (1993)) using codon frequency matrices derived from 122 *H. influenaze* coding sequences in GenBank. The predicted coding region sequences (plus 300 bp of flanking sequence) were used in searches against a database of non-redundant bacterial proteins (NRBP) created specifically for the annotation. Redundancy was removed from NRBP at two stages. All DNA coding sequences were extracted from GenBank (release 85), and sequences from the same species were searched against each other. Sequences having >97% similarity over regions >100 nucleotides were combined. In addition, the sequences w re translated and used in protein comparisons with all sequences in Swiss-Prot

(release 30). Sequences belonging to the same species and having >98% similarity over 33 amino acids were combined. NRBP is composed of 21,445 sequences extracted from 23,751 GenBank sequences and 11,183 Swiss-Prot sequences from 1,099 different species.

A total of 1,749 predicted coding regions were identified. Searches of the *H. influenzae* predicted coding regions were performed using an algorithm that translates the query DNA sequence in the three plus-strand reading frames for searching against NRBP, identifies the protein sequences that match the query, and aligns the protein-protein matches using **praze**, a modified Smith-Waterman (Pearson and Lipman, *Proc. Natl. Acad. Sci. U.S.A. 85:*2444 (1988)) algorithm. In cases where insertion or deletions in the DNA sequence produced a frameshift error, the alignment algorithm started with protein regions of maximum similarity and extended the alignment to the same database match in alternative frames using the 300 bp flanking region. Regions known to contain frameswft errors were saved in the database and evaluated for possible correction. Unidentified predicted coding regions and the remaining intergenic sequences were searched against a dataset of all available peptide sequences from Swiss-Prot, PIR, and GenBank. Identification of operon structures will be facilitated by experimental determination of transcription promoter and termination sites.

Each putatively identified *H. influenaze* gene was assigned to one of 102 biological role categories adapted from Riley (Riley, M., *Microbiology Reviews 57(4)*:862 (1993)). Assignments were made by linking the protein sequence of the predicted coding regions with the Swiss-Prot sequences in the Riley database. Of the 1,749 predicted coding regions, 724 have no role assignment. Of these, no database match was found for 384, while 340 matched "hypothetical proteins" in the database. Role assignments were made for 1,025 of the predicted coding regions. A compilation of all the predicted coding regions, their unique identifiers, a three letter gene identifier, percent id ntity, percent similarity, and amino acid match length are presented in Table 1(a).

An annotated complete genome map of *H. influenaze* Rd is presented in Figures 6(A)-(D). The map places each predicted coding region on the *H. influenaze* chromosome, indicates its direction of transcription and color codes its role assignment. Role assignments are also represented in Figure 5.

A survey of the genes and their chromosomal organization in *H. influenaze* Rd make possible a description of the metabolic processes *H. influenaze* requires for survival as a free living organism, the nutritional requirements for its growth in the laboratory, and the characteristics which make it unique from other organisms specifically as it relates to its pathogenicity and virulence. The genome would be expected to have complete complements of certain classes of genes known to be essential for life. For example, there is a one-to-one correspondence of published *E. coli* ribosomal protein sequences to potential homologs in the *H. influenaze* database. Likewise, as shown in Table 1(a), an aminoacyl tRNA-synthetase is present in the genome for each amino acid. Finally, the location of tRNA genes was mapped onto the genome. There are 54 identified tRNA genes, including representatives of all 20 amino acids.

In order to survive as a free living organism, *H. influenaze* must produce energy in the form of ATP via fermentation and/or electron transport. As a facultative anaerobe, *H. influenaze* Rd is known to ferment glucose, fructose, galactose, ribose, xylose and fucose (Dorocicz *et al.*, *J. Bacteriol. 175:*7142 (1993)). The genes identified in Table 1(a) indicate that transport systems are available for the uptake of these sugars via the phosphoenolpyruvate-phosphotransferase system (PTS), and via non-PTS mechanisms. Genes that specify the common phosphate-carriers Enzyme I and Hpr (*ptsI* and *ptsH*) of the PTS system were identified as well as the glucose specific *crr* gene. The *ptsH*, *ptsI*, and *crr* genes constitute the *pts* operon. We have not however identified the gene encoding membrane-bound glucose specific Enzyme II. The latter enzyme is required for transport of glucose by the PTS system. A complete PTS system for fructose was identified.

Genes encoding the complete glycolytic pathway and for the production of fermentative end products were identified. Growth utilizing anaerobic respiratory mechanisms were found by identifying genes encoding functional electron transport systems using inorganic electron acceptors such as nitrates, nitrites, and dimethylsulfoxide. Genes encoding three enzymes of the tricarboxylic acid (TCA) cycle appear to be absent from the genome. Citrate synthase, isocitrate dehydrogenase, and acordtase were not found by searching the predicted coding regions or by using the *E. coli* enzymes as peptide queries against the entire genome in translation. This provides an explanation for the very high level of glutamate (lg/L) which is required in defined culture media (Klein and Luginbuhl, *J. Gen. Microbiol. 113:*409 (1979)). Glutamate can be directed into the TCA cycle via conversion to alpha-ketoglutarate by glutamate dehydrogenase. In the absence of a complete TCA cycle, glutamate presumably serves as the source of carbon for biosynthesis of amino acids using precursors which branch from the TCA cycle. Functional electron transport systems are available for the production of ATP using oxygen as a terminal electron acceptor.

Previously unanswered questions regarding pathogenicity and virulence can be addressed by examining certain classes of genes such as adhesions and the lipooligosaccharide biogenesis genes. Moxon and co-workers (Weiser *et al., Cell 59:*657 (1989)) have obtained evidence that a number of these virulence-related genes contain tandem tetramer repeats which undergo frequent addition and deletion of one or more repeat units during replication such that the reading frame of the gene is changed and its expression thereby altered. It is now possible, using the complete genome sequence, to locate all such tandem repeat tracts (Figure 5) and to begin to determine their roles in phase variation of such potential virulence genes.

H. influenzae Rd possesses a highly efficient natural DNA transformation system (Kahn and Smith, *J. Membrane Biol. 138:*155 (1984). A unique DNA uptake sequence site, 5′ AAGTGCGGT, present in multiple copies in the genome, has been shown to be necessary for efficient DNA

BNSDOCID: <WO___9633276A1_I_>

uptake. It is now possible to locate all f these sites and completely describe their distribution with respect to genic and intergenic regions. Fifteen genes involved in transformation have already been described and sequenced (Redfield, R., *J. Bacteriol. 173*:5612 (1991); Chandler, M., *Proc. Natl. Acad. Sci. U.S.A 89*:1616 (1992); Barouki and Smith, *J. Bacteriol. 163(2)*:629 (1985); Tomb *et al., Gene 104*:1 (1991); Tomb, J, *Proc. Natl. Acad. Sci. U.S.A 89*:10252 (1992)). Six of the genes, *comA* to *comF*, comprise an operon which is under positive control by a 22-bp palindromic competence regulatory element (CRE) about one helix turn upstream of the promoter. The rec-2 transformation gene is also controlled by this element. It is now possible to locate additional copies of CRE in the genome and discover potential transformation genes under CRE control. In addition, it may now be possible to discover other global regulatory elements with an ease not previously possible.

One well-described gene regulatory system in bacteria is the "two-component" system composed of a sensor molecule that detects some sort of environmental signal and a regulator molecule that is phosphorylated by the activated form of the sensor. The regulator protein is generally a transcription factor which, when activated by the sensor, turns on or off expression of a specific set of genes (for review, see Albright *et al., Ann. Rev. Genet. 23*:311 (1989); Parkinson and Kofoid, *Ann. Rev. Genet. 26*:71 (1992)). It has been estimated that *E. coli* harbors 40 sensor-regulator pairs (Albright *et al., Ann. Rev. Genet. 23*:311 (1989); Parkinson and Kofoid, *Ann. Rev. Genet. 26*:71 (1992)). The *H. influenaze* genome was searched with representative proteins from each family of sensor and regulator proteins using **tblastn** and **tfasta**. Four sensor and five regulator proteins were identified with similarity to proteins from other species (Table 6). There appears to be a corresponding sensor for each regulator protein except CpxR. Searches with the CpxA protein from *E. coli* identified three of the four sensors listed in Table 6, but no additional significant matches were found. It is possible that the level of sequence similarity is low enough to be undetectable with **tfasta**. No

BNSDOCID: <WO__ 9633276A1_I_>

representatives f the NtrC-class of regulators were found. This class of proteins interacts directly with the sigma-54 subunit of RNA polymerase, which is not present in *H. influenaze*. All of the regulator proteins fall into the OmpR subclass (Albright *et al.*, *Ann. Rev. Genet. 23*:311 (1989); Parkinson and Kofoid, *Ann. Rev. Genet. 26*:71 (1992)). The *phoBR* and *basRS* genes of *H. influenaze* are adjacent to one another and presumably form an operon. The *nar* and *arc* genes are not located adjacent to one another.

Some of the most interesting questions that can be answered by a complete genome sequence relate to what genes or pathways are absent. The non-pathogenic *H. influenaze* Rd strain varies significantly from the pathogenic serotype b strains. Many of the differences between these two strains appear in factors affecting infectivity. For example, the eight genes which make up the fimbrial gene cluster (vanHam *et al.*, *Mol. Microbiol. 13*:673 (1994)) involved in adhesion of bacteria to host cells are now shown to be absent in the Rd strain. The *pepN* and *purE* genes which flank the fimbrial cluster in *H. influenaze* type b strains are adjacent to one another in the Rd strain (Fig. 7), suggesting that the entire fimbrial duster was excised. On a broader level, we determined which *E. coli* proteins are not in *H. influenzae* by taking advantage of a non-redundant set of protein coding genes from *E. coli*, namely the University of Wisconsin Genome Project contigs in GenBank: 1,216 predicted protein sequences from GenBank accessions D10483, L10328, U00006, U00039, U14003, and U18997 (Yura *et al.*, *Nucleic Acids Research 20*:3305 (1992); Burland *et al.*, *Genomics 16*:551 (1993)). The minimum threshold for matches was set so that even weak matches would be scored as positive, thereby giving a minimal estimate of the *E. coli* genes not present in *H. influenaze*. **tblastn** was used to search each of the *E. coli* proteins against the complete genome. All **blast** scores >100 were considered matches. Altogether 627 *E. coli* proteins matched at least one region of the *H. influenaze* genome and 589 proteins did not. The 589 non-matching proteins were examined and found to contain a disproportionate number of hypothetical proteins from *E. coli*. Sixty-eight percent of the identified *E. coli* proteins

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

were matched by an *H. influenaze* sequence whereas only 38% of the hypothetical proteins were matched. Proteins are annotated as hypothetical based on a lack of matches with any other known protein (Yura *et al.*, *Nucleic Acids Research 20*:3305 (1992); Burland *et al.*, *Genomics 16*:551 (1993)). At least two potential explanations can be offered for the over representation of hypothetical proteins among those without matches: some of the hypothetical proteins are not, in fact, translated (at least in the annotated frame), or these are *E. coli*-specific proteins that are unlikely to be found in any species except those most closely related to *E. coli*, for example *Salmonella typhimurium*.

A total of 384 predicted coding regions did not display significant similarity with a six-frame translation of GenBank release 87. These unidentified coding regions were compared to one another with fasta. Several novel gene families were identified. For example, two predicted coding regions without database matches (HI0591, HI0852) share 75% identity over almost their entire lengths (139 and 143 amino acid residues respectively). Their similarity to each other but failure to match any protein available in the current databases suggest that they could represent a novel cellular function.

Other types of analyses can be applied to the unidentified coding regions, including hydropathy analysis, which indicates the patterns of potential membrane-spanning domains that are often conserved between members of receptor and transporter gene families, even in the absence of significant amino acid identity. Five examples of unidentified predicted coding regions that display potential transmembrane domains with a periodic pattern that is characteristic of membrane-bound channel proteins are shown in Figure 8. Such information can be used to focus on specific aspects of cellular function that are affected by targeted deletion or mutation of these genes.

Interest in the medically important aspects of *H. influenaze* biology has focused particularly on those genes which determine virulence characteristics of the organism. Recently, the catalase gene was characterized and sequenced as a possible virulence-related gene (Bishai *et al.*, *J. Bacteriol. 176*:2914

BNSDOCID: <WO 9633276A1 I >

(1994)). A number of the genes responsible for the capsular polysaccharide have been mapped and sequenced (Kroll *et al., Mol. Microbiol. 5(6)*:1549 (1991)). Several outer membrane protein genes have been identified and sequenced (Langford *et al., J. Gen. Microbiol. 138*:155 (1992)). The lipooligosaccharide component of the outer membrane and the genes of its synthetic pathway are under intensive study (Weiser *et al., J. Bacteriol. 173*:3304 (1990)). While a vaccine is available, the study of outer membrane components is motivated to some extent by the need for improved vaccines.

Data Availability

The *H. influenaze* genome sequence has been deposited in the Genome Sequence DataBase (GSDB) with the accession number L42023. The nucleotide sequence and peptide translation of each predicted coding region with identified start and stop codons have also been accessioned by GSDB.

Production of an Antibody to a Haemophilus influenzae Protein

Substantially pure protein or polypeptide is isolated from the transfected or transformed cells using any one of the methods known in the art. The protein can also be produced in a recombinant prokaryotic expression system, such as *E. coli*, or can by chemically synthesized. Concentration of protein in the final preparation is adjusted, for example, by concentration on an Amicon filter device, to the level of a few micrograms/ml. Monoclonal or polyclonal antibody to the protein can then be prepared as follows:

Monoclonal Antibody Production by Hybridoma Fusion

Monoclonal antibody to epitopes of any of the peptides identified and isolated as described can be prepared from murine hybridomas according t the classical method f Kohler, G. and Milstein, C., *Nature 256*:495 (1975) r

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

modifications of the methods thereof. Briefly, a mouse is repetitively inoculated with a few micrograms f the selected protein over a period of a few weeks. The mouse is then sacrificed, and the antibody producing cells of the spleen isolated. The spleen cells are fused by means of polyethylene glycol with mouse myeloma cells, and the excess unfused cells destroyed by growth of the system on selective media comprising aminopterin (HAT media). The successfully fused cells are diluted and aliquots of the dilution placed in wells of a microtiter plate where growth of the culture is continued. Antibody-producing clones are identified by detection of antibody in the supernatant fluid of the wells by immunoassay procedures, such as ELISA, as originally described by Engvall, E., *Meth. Enzymol. 70*:419 (1980), and modified methods thereof. Selected positive clones can be expanded and their monoclonal antibody product harvested for use. Detailed procedures for monoclonal antibody production are described in Davis, L. *et al.* Basic Methods in Molecular Biology Elsevier, New York. Section 21-2 (1989).

Polyclonal Antibody Production by Immunization

Polyclonal antiserum containing antibodies to heterogenous epitopes of a single protein can be prepared by immunizing suitable animals with the expressed protein described above, which can be unmodified or modified to enhance immunogenicity. Effective polyclonal antibody production is affected by many factors related both to the antigen and the host species. For example, small molecules tend to be less immunogenic than other and may require the use of carriers and adjuvant. Also, host animals vary in response to site of inoculations and dose, with both inadequate or excessive doses of antigen resulting in low titer antisera. Small doses (ng level) of antigen administered at multiple intradermal sites appears to be most reliable. An effective immunization protocol for rabbits can be found in Vaitukaitis, J. *et al.*, *J. Clin. Endocrinol. Metab. 33*:988-991 (1971).

Booster injections can be given at regular intervals, and antiserum harvested when antibody titer thereof, as determined semi-quantitatively, for example, by double immunodiffusion in agar against known concentrations of the antigen, begins to fall. See, for example, Ouchterlony, O. *et al.*, Chap. 19 in: *Handbook of Experimental Immunology*, Wier, D., ed, Blackwell (1973). Plateau concentration of antibody is usually in the range of 0.1 to 0.2 mg/ml of serum (about 12 μM). Affinity of the antisera for the antigen is determined by preparing competitive binding curves, as described, for example, by Fisher, D., Chap. 42 in: *Manual of Clinical Immunology*, second edition, Rose and Friedman, eds., Amer. Soc. For Microbiology, Washington, D.C. (1980).

Antibody preparations prepared according to either protocol are useful in quantitative immunoassays which determine concentrations of antigen-bearing substances in biological samples; they are also used semi-quantitatively or qualitatively to identify the presence of antigen in a biological sample.

Preparation of PCR Primers and Amplification of DNA

Various fragments of the *Haemophilus influenzae* Rd genome, such as those disclosed in Tables 1(a) and 2 can be used, in accordance with the present invention, to prepare PCR primers for a variety of uses. The PCR primers are preferably at least 15 bases, and more preferably at least 18 bases in length. When selecting a primer sequence, it is preferred that the primer pairs have approximately the same G/C ratio, so that melting temperatures are approximately the same. The PCR primers and amplified DNA of this Example find use in the Examples that follow.

Gene expression from DNA Sequences Corresponding to ORFs

A fragment of the *Haemophilus influenzae* Rd genome provided in Tables 1(a) or 2 is introduced into an expression vector using conventional

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO___9633276A1_I_>

technology. (Techniques to transfer cloned sequences into expression vectors that direct protein translati n in mammalian, yeast, insect or bacterial expression systems are well known in the art.) Commercially available vectors and expression systems are available from a variety of suppliers including Stratagene (La Jolla, California), Promega (Madison, Wisconsin), and Invitrogen (San Diego, California). If desired, to enhance expression and facilitate proper protein folding, the codon context and codon pairing of the sequence may be optimized for the particular expression organism, as explained by Hatfield *et al.*, U.S. Patent No. 5,082,767, incorporated herein by this reference.

The following is provided as one exemplary method to generate polypeptide(s) from cloned ORFs of the *Haemophilus* genome fragment. Since the ORF lacks a poly A sequence because of the bacterial origin of the ORF, this sequence can be added to the construct by, for example, splicing out the poly A sequence from pSG5 (Stratagene) using *BglI* and *SalI* restriction endonuclease enzymes and incorporating it into the mammalian expression vector pXT1 (Stratagene) for use in eukaryotic expression systems. pXT1 contains the LTRs and a portion of the gag gene from Moloney Murine Leukemia Virus. The position of the LTRs in the construct allow efficient stable transfection. The vector includes the Herpes Simplex thymidine kinase promoter and the selectable neomycin gene. The *Haemophilus* DNA is obtained by PCR from the bacterial vector using oligonucleotide primers complementary to the *Haemophilus* DNA and containing restriction endonuclease sequences for *PstI* incorporated into the 5′ primer and *BglII* at the 5′ end of the corresponding *Haemophilus* DNA 3′ primer, taking care to ensure that the *Haemophilus* DNA is positioned such that its followed with the poly A sequence. The purified fragment obtained from the resulting PCR reaction is digested with *PstI*, blunt ended with an exonuclease, digested with *BglII*, purified and ligated to pXT1, now containing a poly A sequence and digested *BglII*.

SUBSTITUTE SHEET (RULE 26)

BNSDOCID: <WO__9633276A1 I >

The ligated product is transfected into mouse NIH 3T3 cells using Lipofectin (Life Techn logies, Inc., Grand Island, New York) under conditions outlined in the product specification. Positive transfectants are selected after growing the transfected cells in 600 ug/ml G418 (Sigma, St. Louis, Missouri). The protein is preferably released into the supernatant. However if the protein has membrane binding domains, the protein may additionally be retained within the cell or expression may be restricted to the cell surface.

Since it may be necessary to purify and locate the transfected product, synthetic 15-mer peptides synthesized from the predicted *Haemophilus* DNA sequence are injected into mice to generate antibody to the polypeptide encoded by the *Haemophilus* DNA.

If antibody production is not possible, the *Haemophilus* DNA sequence is additionally incorporated into eukaryotic expression vectors and expressed as a chimeric with, for example, ß-globin. Antibody to ß-globin is used to purify the chimeric. Corresponding protease cleavage sites engineered between the ß-globin gene and the *Haemophilus* DNA are then used to separate the two polypeptide fragments from one another after translation. One useful expression vector for generating ß-globin chimerics is pSG5 (Stratagene). This vector encodes rabbit ß-globin. Intron II of the rabbit ß-globin gene facilitates splicing of the expressed transcript, and the polyadenylation signal incorporated into the construct increases the level of expression. These techniques as described are well known to those skilled in the art of molecular biology. Standard methods are published in methods texts such as Davis et al. and many of the methods are available from the technical assistance representatives from Stratagene, Life Technologies, Inc., or Promega. Polypeptide may additionally be produced from either construct using in vitro translation systems such as In vitro Express™ Translation Kit (Stratagene).

While the present invention has been described in some detail for purposes of clarity and understanding, one skilled in the art will appreciate that various

changes in form and detail can be made without departing from the true scope of the invention.

All patents, patent applications and publications referred to above are hereby incorporated by reference.

Amino acid biosynthesis							
	Glutamate family						
HI0190	202698	204044	glutamate dehydrogenase (gdhA) (Escherichia coli)	74.1	84.4	446	
HI0867	915793	917833	glutamine synthetase (glnA) (Proteus vulgaris)	70.7	85.9	467	
HI1725	1792409	1789821	uridylyl transferase (glnD) (Escherichia coli)	46.6	67.8	854	
HI0813	861610	860240	argininosuccinate lyase (arginosuccinase) (asaI) (argH) (Escherichia coli)	73.5	84.5	457	
HI1733	1799112	1800443	argininosuccinate synthetase (argG) (Escherichia coli)	78.6	87.5	438	
HI0598	618753	617752	ornithine carbamoyltransferase (arcB) (Pseudomonas aeruginosa)	82.3	90.7	334	
HI1242	1313013	1311763	gamma-glutamyl phosphate reductase (proA) (Escherichia coli)	61.7	79.4	406	
HI0902	955518	956621	glutamate 5-kinase (gamma-glutamyl kinase) (proB) (Escherichia coli)	65.7	80.2	363	

Table 1(a)

76.1

BNSDOCID: <WO___9633276A1_I_>

HI0215	229004	231046	oligopeptidase A (prlC) (Escherichia coli)	72.0	84.8	678
HI0677	716670	718121	peptidase D (pepD) (Escherichia coli)	56.8	72.2	485
HI0589	608542	607865	peptidase E (pepE) (Escherichia coli)	41.4	60.0	214
HI1351	1423832	1425067	peptidase T (pepT) (Salmonella typhimurium)	53.3	71.4	398
HI1262	1336467	1335070	periplasmic serine protease Do and heat shock protein (htrA) (Escherichia coli)	55.8	73.9	469
HI1603	1664636	1663212	probable ATP-dependent protease (sms) (Escherichia coli)	80.0	92.2	460
HI0724	768169	768786	proline dipeptidase (pepQ) (Escherichia coli)	53.7	70.2	204
HI0137	151209	151901	protease (prtH) (Porphyromonas gingivalis)	52.6	64.9	57
HI1547	1613228	1611384	protease IV (sppA) (Escherichia coli)	43.7	64.0	607
HI0152	167927	166698	protease specific for phage lambda cII repressor (hflK) (Escherichia coli)	55.8	72.6	396
HI1688	1751031	1752089	putative protease (sohB) (Escherichia coli)	53.3	74.5	348
HI0532	553214	552189	sialoglycoprotease (gcp) (Pasteurella haemolytica)	81.8	91.5	319
Transport/binding proteins						
	Amino acids, peptides, amines					
HI1183	1247387	1246659	arginine transport ATP-binding protein artP (artP) (Escherichia coli)	65.8	83.1	242
HI1180	1245250	1244570	arginine transport system permease protein (artM) (Escherichia coli)	55.7	79.9	218
HI1181	1245915	1245253	arginine transport system permease protein (artQ) (Escherichia coli)	59.0	77.8	229
HI0254	284235	283786	biopolymer transport protein (exbB) (Haemophilus influenzae)	96.0	98.7	150
HI0253	283779	283339	biopolymer transport protein (exbD) (Escherichia coli)	28.8	55.1	118
HI1734	1801710	1800520	branched chain aa transport system II carrier protein (braB) (Pseudomonas aeruginosa)	28.4	49.8	279
HI0885	935516	934149	D-alanine permease (dagA) (Alteromonas haloplanktis)	43.2	65.5	527
HI1188	1251117	1250128	dipeptide transport ATP-binding protein (dppD) (Escherichia coli)	74.2	84.0	326
HI1187	1250122	1249142	dipeptide transport ATP-binding protein (dppF) (Escherichia coli)	76.4	87.1	325
HI1126	1189626	1188709	dipeptide transport system permease protein (dppB) (Escherichia coli)	34.1	60.7	337

SUBSTITUTE SHEET (RULE 26)

HI1190	1253029	1252031	dipeptide transport system permease protein (dppB) (Escherichia coli)	61.1	79.2	337
HI1189	1252013	1251130	dipeptide transport system permease protein (dppC) (Escherichia coli)	63.8	83.3	287
HI1536	1601926	1603137	glutamate permease (gltS) (Escherichia coli)	53.9	73.0	391
HI1081	1146102	1145389	glutamine transport system permease protein (glnP) (Escherichia coli)	37.6	59.0	212
HI1082	1146859	1146089	glutamine-binding periplasmic protein (glnH) (Escherichia coli)	28.4	48.2	222
HI0410	429066	428263	leucine-specific transport protein (livG) (Escherichia coli)	28.1	55.2	250
HI0227	255068	256375	membrane-associated component, LIV-II transport system (bmQ) (Salmonella typhimurium)	32.9	60.4	425
HI0214	228528	226987	oligopeptide binding protein (oppA) (Escherichia coli)	31.7	53.5	473
HI1127	1191333	1189710	oligopeptide binding protein (oppA) (Escherichia coli)	52.6	69.0	527
HI1124	1187751	1186783	oligopeptide transport ATP-binding protein (oppD) (Salmonella typhimurium)	77.2	85.0	320
HI1123	1186783	1185788	oligopeptide transport ATP-binding protein (oppF) (Salmonella typhimurium)	71.5	83.9	329
HI1125	1188696	1187764	oligopeptide transport system permease protein (oppC)C (Salmonella typhimurium)	71.1	87.4	300
HI1644	1702355	1704049	peptide transport periplasmic protein (sapA) (Salmonella typhimurium)	39.3	63.8	504
HI1647	1705898	1706944	peptide transport system ATP-binding protein (sapD) (Salmonella typhimurium)	62.4	80.0	330
HI1646	1705007	1705891	dipeptide transport system permease protein (dppC) (Escherichia coli)	36.2	59.9	279
HI1645	1704052	1705014	peptide transport system permease protein (sapB) (Salmonella typhimurium)	34.4	63.8	319
HI1182	1246638	1245922	periplasmic arginine-binding protein (artI) (Pasteurella haemolytica)	58.6	73.4	234
HI1157	1221270	1222589	proton glutamate sympot protein (gltP) (Bacillus caldotenax)	26.6	53.6	395
HI0592	611920	610616	putrescine transport protein (potE) (Escherichia coli)	77.2	88.0	434
HI0291	324543	323308	serine transporter (sdaC) (Escherichia coli)	61.0	77.8	411

OCID: <WO___9633276A1_I_>

HI0129	143015	144800	nitrogenase C (nifC) (Clostridium pasteurianum)	27.1	52.6	248	
HI1480	1559124	1558768	nitrogenase C (nifC) (Clostridium pasteurianum)	40.9	60.2	92	
HI0359	381523	382464	nmt1 protein (nmt1) (Aspergillus parasiticus)	25.6	54.7	289	
HI1299	1375415	1374882	partitioning system protein (parB) (Plasmid RP4)	43.6	67.7	141	
HI0224	252941	252168	rarD protein (rarD) (Escherichia coli)	26.5	53.0	230	
HI0682	721733	720840	rarD protein (rarD) (Escherichia coli)	27.1	55.0	289	
HI0918	970839	970249	skp protein (skp) (Pasteurella multocida)	55.5	76.4	191	
HI0983	1038375	1037893	small protein (smpB) (Escherichia coli)	78.8	91.3	160	
HI1598	1661468	1659882	spoIIIE protein (spoIIIE) (Coxiella burnetii)	56.1	74.5	504	
HI0898	951407	952018	suppressor protein (msgA) (Escherichia coli)	30.2	56.1	254	
HI1080	1145382	1144612	surfactin (sfpo) (Bacillus subtilis)	58.2	77.9	246	
HI0753	811790	811296	toxR regulon (tagD) (Vibrio cholerae)	45.7	64.0	164	
HI1412	1502860	1501311	traN protein (traN) (Plasmid RP4)	40.2	61.5	233	
HI0666	708305	709960	transport ATP-binding protein (cydC) (Escherichia coli)	26.3	51.7	536	
HI1159	1225137	1223410	transport ATP-binding protein (cydC) (Escherichia coli)	48.5	70.1	568	
HI1562	1627239	1626295	vanH protein (vanH) (Transposon Tn1546)	39.7	57.1	251	
HI0632	668489	669433	mucoid status locus protein (mucB) (Pseudomonas aeruginosa)	25.4	51.8	309	
HI0172	183553	184785	phenolhydroxylase (ORF6) (Acinetobacter calcoacelicus)	33.0	56.9	313	
HI1390	1481177	1481266	plasma protease C1 inhibitor (Homo sapiens)	75.0	79.2	23	

(1) GENERAL INFORMATION:

(i) APPLICANTS: Human Genome Sciences, Inc.;
Johns Hopkins University

(ii) TITLE OF INVENTION: The Nucleotide Sequence of the Haemophilus influenzae Rd Genome, Fragments Thereof, and Uses Thereof

(iii) NUMBER OF SEQUENCES: 1

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: Sterne, Kessler, Goldstein & Fox, P.L.L.C.

(B) STREET: 1100 New York Avenue, N.W.

(C) CITY: Washington

(D) STATE: DC

(E) COUNTRY: USA

(F) ZIP: 20005-3934

(v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: 3 1/2 inch diskette

(B) COMPUTER: Dell Pentium

(C) OPERATING SYSTEM: MS DOS v6.22

(D) SOFTWARE: ASCII Text

(vi) CURRENT APPLICATION DATA:

(A) APPLICATION NUMBER: (Not Yet Assigned)

(B) FILING DATE: (Herewith)

(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: US 08/476,102

(B) FILING DATE: 07-JUN-1995

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: US 08/487,429

(B) FILING DATE:07-JUN-1995

AGAACTTGCT GGTTTACCTG AATCCGCACT ACAAGCTGCA CAACAATCTG CCGAAAGTAA 229620

AGGATTAAAG GGCTATCGTT TTACGTTAGA AATCCCAAGT TATTTGCCTG TGATGACTTA 229680

CTGCGAAAAT CGTGCATTGC GTGAAGAAAT GTACCGTGCT TATGCAACAC GTGCCTCTGA 229740

ACAAGGCCCT AACGCTGGCA AATGGGACAA CAGTAAAGTT ATGGAAGAAA TTCTGACGTT 229800

ACGCGTTGAA CTGGCAAAAT TACTCGGTTT TAATACTTAT ACCGAACTTT CACTTGCCAC 229860

AAAAATGGCA GAAAACCCAC AACAAGTGCT TGATTTCTTA GATCATCTTG CCGAACGTGC 229920

CAAACCACAA GGCGAAAAAG AACTGCAAGA ATTGAAAGGT TACTGCGAAA AAGAATTTGG 229980

TGTCACTGAA CTTGCACCTT GGGATATTGG TTTTTACAGC GAAAAACAAA AACAACATTT 230040

ATACGCCATT AATGATGAAG AACTTCGCCC ATATTTTCCA GAAAATCGCG TGATTTCAGG 230100

ATTGTTTGAA TTAATTAAAC GCATTTTTAA TATTCGTGCA GTGGAACGCA AAGGTGTGGA 230160

TACTTGGCAT AAAGATGTAC GTTTCTTTGA TTTAATCGAT GAAAATGATC AACTCCGTGG 230220

CAGTTTCTAT CTTGATTTAT ATGCACGCGA ACACAAACGT GGTGGTGCGT GGATGGATGA 230280

CTGTATCGGA CGTAAGCGTA AACTGGATGG CAGCATTGAA ACGCCTGTGG CTTATTTAAC 230340

CTGTAACTTC AACGCACCGA TAGGGAATAA ACCAGCGTTA TTTACTCACA ATGAAGTCAC 230400

CACGCTATTC CACGAATTTG GGCACGGAAT CCATCATATG CTCACGCAAA TTGATGTATC 230460

CGATGTGGCG GGAATAAATG GCGTGCCTTG GGATGCTGTG GAATTACCAA GTCAATTTAT 230520

GGAAAACTGG TGCTGGGAAG AAGAGGCTTT GGCGTTTATT TCAGGACATT ACGAAACAGG 230580

CGAGCCGTTG CCAAAAGAAA AACTCACGCA ATTGTTAAAA GCGAAAAATT TCCAAGCTGC 230640

GATGTTTATC TTACGTCAGC TTGAATTTGG TATTTTTGAT TTCCGTTTAC ATCACACCTT 230700

TGATGCGGAA AAAACCAATC AAATTTTGGA TACACTTAAA TCAGTGAAAT CACAAGTTGC 230760

CGTCATAAAA GGCGTAGATT GGGCAAGAGC GCCACACAGC TTCAGCCATA TTTTTGCAGG 230820

GGGCTACGCC GCAGGTTATT ACAGCTATTT ATGGGCTGAA GTATTATCGG CTGATGCTTA 230880

TTCACGCTTT GAAGAAGAAG GTATTTTCAA CCCAATCACC GGCAAATCAT TCTTAGATGA 230940

AATTCTCACC CGTGGCGGCT CAGAAGAACC AATGGAACTC TTCAAACGCT TCCGTGGCCG 231000

TGAACCGCAA TTAGATGCGT TGTTAAGACA TAAAGGAATT ATGAACTAAT ACGTCAAACT 231060

ATAGGGCATG CACAACATGC CCTTTCTCAA ATATTTTCCT CGACATAAAA ACATAGTAAA 231120

TTTTAACCGC ACTTTATCTA GACAAAACCT GCAAACGAGA CATTATTTTT ACAAATAGCT 231180

GTAAGATAAA CGCATAGAAA TTCACGCAGA ATAAACTAAT TTCCTGTATA ATTTTTCACA 231240

3NSDOCID: <WO___9633276A1_I_>

TATCTATCCA ACTCATTACA CAAAATATTC TGAGGATTTT ATGGCTGCTG CAATTCAACA 231300

ACGTGCCGAA CTTCAACGCC GTATTTGGCA AATTGCTAAT GATGTGCGAG GCTCGGTCGA 231360

TGGCTGGGAT TTCAAACAAT ATGTGCTTGG CACACTTTTT TACCGTTTTA TTAGCGAAAA 231420

TTTTGCCAAT TACATTGAAG CGGGCGATGA AAGCGTAAAT TATGCCCAAT TACCTGATGA 231480

AATCATTACA CAGATGCCAT TAAAACGAAA GGCTACTTTA TTTACCCAAG CCAATTATTT 231540

AAGAATGTTG CGGCTAATGC TGGCAGCAAT CCTAATTTGA ATACTGATTT AAAACAGATT 231600

TTTACTGATA TTGAAAACTC AGCGACGGGC TTTCCGTCTG AACAAGATAT TAAAGGGTTA 231660

TTTGCCGATT TTGATACCAC CAGCAATCGC TTAGGCAATA CCGTAAAAGA TAAAAACGAC 231720

CGCTTAACGG CTGTTTTGAA AGGCGTGGCT GAACTTGATT TTGGCAAATT TGAAGATAAC 231780

CACATTGATT TATTTGGCGA TGCATACGAA TATCTTATTT CTAACTATGC CGCCAATGCA 231840

GGCAAATCTG GTGGCGAATT TTTTACCCCA CAAAGTGTTT CCAAACTCAT TGCTCAAATT 231900

GCAATGCACG GGCAAACCTC GGTCAATAAA ATTTATGACC CTGCAGCAGG TTCTGGCTCA 231960

CTTTTGCTTC AAGCCAAAAA ACAATTTGAT GAACATATTA TTGAAGAAGG CTTTTTCGGG 232020

CAGGAAATTA ACCATACCAC ATACAACCTT GCCCGTATGA ATATGTTTTT GCATAACATC 232080

AACTACGACA AGTTTGATAT TGCTTTAGGC AACACCTTAA TGGAACCACA ATTTGGCGAT 232140

AATAAACCTT TCGATGCCAT TGTTTCGAAC CCGCCTTACT CCGTGAAATG GGCTGGCTCC 232200

GACGATCCAA CATTGATTAA TGATGAACGA TTTGCCCCCC GCAGGCGTGC TTGCACCAAA 232260

ATCCAAAGCG GACTTTGCCT TTATTTTACA TGCGTTAAGT TATCTTTCAG CAAAAGGCCG 232320

CGCGGCGATT GTTTCCTTCC CTGGTATTTT TTATCGTGGC GGTGCCGAGC AAAAAATTCG 232380

TCAATATTTG GTGGATAATA ACTATGTGGA CGCGGTGATT GCGCTTGCGC CAAATCTCTT 232440

TTTTGGCACC AGTATTGCGG TGAATATTTT GGTGCTTTCC AAACACAAAC CCAATTTATC 232500

GATGCCAGCG GTTTATTTAA ATCTGCCACT AATAACCACA TTTTAGAAGA GGAACATATC 232560

GAGCAAATCC TCAAACTGTT TGCCGATAAA GAGGATGTGC CGCATTTGGC AAAATCCATA 232620

TCCTTTGAAG AAATCGCCCA AAATGACTAC AACCTTGCAG TAAGTTCGTA TGTGGAACAA 232680

AAAGACACTC GTGAAGTGAT TAATATTGAT GAACTCAATG CTCAAATTCG TGaAACTGTT 232740

ACCAATATTG ACCACTTGCG TGCGGAAATT GACAAGATTG TTGCAGAAAT TGAAGGGTAA 232800

AATGTAGGGT GGGCTTCAGC CCAnaTTTAC GATAACAGGT AGTGGGCTAA AGCCCaCCCT 232860

ACGAGGAACA ATAAAAAAAC ACCATGAAAA ACAACCGCAC TTTTTTAGAA AAATTATTGG 232920

GGAAATTATC CCAGGTACAG AAAACTGGCC GCATTTTGTA CGCAACAAAT CTGAACGCTT 815940

TGAAGCGCGT GTGAGTTTAG TGAAAATTAA TGAAGTGGAT TCAGTTTGGT TTGCAGGAAT 816000

GGCTGGCTCA CATATGCCCA TTGCGGTTTC TCACGGGGAA GGACAAGTTA AGTTTAAATC 816060

AGTTGAACAG TTTGCAGGCT TAAAAGCACA AGGCATCATT GCCGCACAAT ATATTGACAA 816120

CAACGGTAGC CCCACCGAAC TATATCCTGC CAATCCGAAT GGTTCTTCGG AGGGCATTAC 816180

TGCAATTACA AATCTTGATG GGCGAGTTGC TATTATGATG CCACACCCTG AACGTGTGTT 816240

CCGTGCTGTG AGTAACTCTT GGCACCCAGA AAACTGGACA GAAGACGGTG CGTGGATGAG 816300

ATTATTCAGA AATGCAAGAA TGGTATTTTA GAGAAATAAA TGGACTTGAA GGTTAAAAAA 816360

GTGCGGTGGA AATCTGGCAG ATTTTCACTG TGCTTCGGTG TTTTCAATTT TATGTAATTG 816420

TTTTTAGTTG TGTTTTTGTT TGATGATTTA ATTATTTTTA TTGTTAAGTT AGCGATATAA 816480

TCATCAAAAT AGAAATATAG AAATGATTAA TATGAAACCT CAATATTATC TTGGTATGAT 816540

GTCGGGTACT AGCTTAGATG GCGTGGATAT TGTGTTGGTT GATTTTTCTC AAGATCCTCA 816600

ACTTATCCTA TCTGATTTTT TTCCTATGCC AGAAGATTTG CGTGAGAAAT TAACCACACT 816660

TATTCAAGTA GGTGAAACAA CTCTACAAAA TTTAGGCGAA CTGGATCATA AACTTGCCTT 816720

GCTATACAGT GATTGTGTCA ATGCTTTTCT GCAGAAAAAT ACTTTTTTGC CAAATCAAAT 816780

TCAGGCTATT GGTTGCCACG GGCAAACGGT ATGGCATTCT CCAAATTCAC AATTTCCCTT 816840

TACAATGCAA TTAGGCGATA TGAATTTACT TGCCGCTAAA ACAGGCATTT CGGTTATTGG 816900

GGATTTTCGC CGTAAGGATA TGGCTTGGGG TGGACAAGGT GCGCCTCTTG TGCCTGCGTT 816960

TCACGAGGCG GTATTTTCAA ATTCTAATTT TGCTACAGCT GTTTTAAATA TTGGCGGGAT 817020

TAGTAATGTG TCTATTTTGT TTCCAAATCA AGCTGTCATC GGCTTTGATA CAGGTCCAGG 817080

TAATACATTA TTAGATCAAT GGATCGAAAA ACATCAAGGG CTTCGTTATG ATGAAAATGG 817140

CGAGTGGGCA GCAAAAGGCA ACGTGAATAA AGTTCTGCTT GATGAATTAC TCAATGAACC 817200

TTTCTTTTCT TTACCTGCAC CCAAGAGTAC TGGGAGAGAA TTATTTAACC TTGTTTGGTT 817260

AAATCACAAA ATAGCAAAAA TAAGGGAAAA ATTGACCGCA CTTTCAGTTG AAATGTCTTT 817320

TCGCCCAGAG GATGTGCAAG CGACGCTTGT TGAATTGACG GTAACAAGCA TTGTTAATGC 817380

ACTAAATCAG CTCCAAACAG ATTTACCAAA ACGTTTACTA GTATGTGGCG GTGGTGCCAA 817440

GAATAGCTTA ATTATGCGTG GATTGCATGA TAATTTACTC GATTGGCAGG TCAGCACAAC 817500

GACAGAGCAA GGTTTTGATA TTGATTATGT CGAAGCCGCT GCTTTTGCTT GGTTAGCTTA 817560

BAD ORIGINAL

BNSDOCID: <WO___9633276A1_I_>

TTGTCGAATC AACAATTTAC CTGCAAATCT ACCAAGTGTA ACTGGGGCAA AAAGTGCGGT 817620

TAGTTTGGGG GCTATTTTTC CTAAAGACTA AGGTCAAAAT ATGAATGACA TTATATTAAA 817680

AAGTTTATCC ACATTAATTA CGGAACAAAG AAATCCGAAT TCAGTTGATA TTGATCGTCA 817740

AAGTACATTA GAAATTGTTC GATTAATGAA TGAAGAAGAT AAACTTGTGC CACTTGCGAT 817800

TGAAAGCTGT TTACCGCAGA TTTCGTTGGC GGTAGAGCAA ATTGTGCAGG CTTTTCAACA 817860

AGGCGGGCGT TTAATTTATA TTGGTGCAGG CACCAGTGGT CGTCTTGGTG TATTGGATGC 817920

GTCAGAATGC CCTCCAACAT TTGGCGTGTC TACGGAAATG GTTAAAGGAA TAATTGCAGG 817980

TGGTGAGTGT GCTATTCGTC ATCCTGTTGA GGGTGCTGAA GATAATACGA AAGCGGTGCT 818040

TAATGATTTA CAAAGCATAC ATTTTTCAAA AAATGATGTA TTAGTTGGTA TCGCTGCAAG 818100

TGGTCGCACA CCTTATGTTA TTGCAGGTCT TCAGTATGCA AAATCGCTTG GTGCATTGAC 818160

AATTTCTATT GCGAGTAATC CAAAATCTGA GATGGCTGAG ATAGCCGATA TTGCGATAGA 818220

GACTATTGTT GGCCCTGAAA TATTGACGGG TTCTAGCCGC TTAAAATCTG GCACAGCACA 818280

AAAAATGGTG CTGAATATGC TGACTACGGC AAGTATGATT TTATTGGGTA AATGTTACGA 818340

AAATTTGATG GTCGATGTAC AAGCCAGTAA TGAAAAGCTT AAAGCCCGTG CGGTGCGGAT 818400

TGTGATGCAA GCGACCGATT GCAATAAAAC ATTAGCAGAA CAGACATTAC TTGAAGCAGA 818460

TCAAAATGCC AAACTGGCGA TTATGATGAT TTTAAGCACT TTATCAAAAT CAGAAGCGAA 818520

AGTATTGTTA GAAAGACATC AAGGTAAATT GAGAAATGCG CTTTCTAAAT AAGACAAAAA 818580

CCTGATTGTA AAATCAGGTT TTTATTTTGG AATTCCACGC AATAATCCTA CAGGCTCAAA 818640

AGGTTCTATA GAAGTTGGTG CAGGCACTTC GCTGAATAAT AAAATAAATG GCTTGGGCGG 818700

GATTACTTTT TCATTACGCC ATAAGTTGCC CATTCCGACT AATTGAATAT CTTCGGGTAA 818760

ATTACGAAGA CCTGCTTGTA AAACAGCAAT CGTATTTGGA CGAGGATGAC CTATTGCAAT 818820

GGCGGAGCCG TGTTTTCTAG CATAGTGAAT TGCTGCTTTA AATTGACGTT GAACGTCTGC 818880

GAATTCGTTA CTATCATCTA AAAATATATG GCGATCTAAT GAACGCACGC CTTGTTCTTT 818940

TGCTATTTTC CCAGCTACAG ATTTTCCTAT TGTTCGACTA TCTAAAAAGA ACAAGTGTTT 819000

TTCTTGAAGT GCGGTCATTA AATAAGTCAT TAATTGGGGA TCTGCCGTAG CTGCACTCCC 819060

CATATGATTG TTCATACCGA TGGCATCACG CACAATATTT TTAGCCGTAT TAACGCGATC 819120

ATTTACTTGT GCCGCAGACA TTCCTAAATG CAGCCCACCG TCTTCTATTT TCACTGCACT 819180

TACTGGTTGC ATCGGCATAT GGATTAAAAT ATCTCTACCT TGAGATTTTG CTTCTTGATT 819240

BAD ORIGINAL

BNSDOCID: <WO___9633276A1_I_>

AAATAAAACC GCACTTTAAA GTGCGGTTTT TTATTGCATC GAATGTTAGA CTAAAGTAAC 1024260

AACATCGATC CCCATTTGAT AATATCAAAG AAGAATTTAT TTTGATTGGT TGCAATGCCA 1024320

TCAGAATTTT TTGGTGTGCC ATCTTCATTT TTTTGCACTT CCACCATGCC ATTTTTATAT 1024380

TGACCTTTCA CGGCAGCCAA ATCATCGCGC GCTTCGTTAC GCAATTTCAA ACAAGACTGT 1024440

TTAGTTAAGG GTTCAACATT GCTCACTCTG GTACGGAATT TTTTAAACTG ATAGCTTGCA 1024500

TAATTCATTG ATTTTTCATC GGCTTGGATA ATCTTCACAT ATTGTTCGCC AATAATTTTT 1024560

TCCAACGGAT AGAAAAAGAC ATATTGTGAG TGAATATAGC TATCTTCTTT TGAAAAATGC 1024620

TGTTGCAAAA TTCGAGTTAA ATTTGGGTAA ACGCATTGTT CTACTTGCTT ACTTGCAATC 1024680

GCCCATTGTT TTGCATGTTG ATCAGACAGT TGATAATCAG CTCCAGCAAA TTCCGCTGGA 1024740

ATTGCGGATT GTTGGCTATC AAACATTGAA CAGCCAGACA GCAATGCTGC CATACCTAAG 1024800

GTTGCAATTA ATTTCACATT TGATCCTTAT AAAAGTGAGA ATGTGCAAAA TTCTAGCAAT 1024860

ATCTTCACAA CAGATAAAGC AGCAAATGAA TTTAGTTTTC CTGATCTGAA AAACTTCCGC 1024920

TATAATGACC GCACTTTTTT ACGTTAATTA AAAGGAATTC TTATGCAAAA TCCAAAAGAT 1024980

GATGTTTTAT ATGCGCCTGT TGAATGGATC GATCACAGTG AAGGCTATAG CGATATTCGC 1025040

TATCATAAAT CCACTGATGG TATTGCAAAA ATTACCATTA ACCGCCCTGA AGTTCGTAAT 1025100

GCGTTCCGTC CACAAACAGT GAAAGAAATG ATGACTGCAT TTTCTGATGC ACGTTTCGAT 1025160

GAAAACATTG GTGTCATTGT ATTAACTGGC GAAGGCGAAA AAGCATTCTG TTCTGGTGGT 1025220

GACCAAAAAG TGCGTGGCGA TTACGGTGGT TACAAAGATG ACAGTGGCGT ACATCACTTA 1025280

AATGTATTAG ATTTCCAACG CGATATTCGT TCTTGTCCAA AACCAGTTGT AGCAATGGTG 1025340

GCTGGTTATG CAATTGGTGG TGGCCATGTA TTACATATGC TTTGCGACTT AACGATTGCT 1025400

GCTGAAAATG CCATTTTCGG GCAAACTGGG CCAAAAGTCG GTTCATTCGA TGGCGGTTGG 1025460

GGCGCAAGCT ATATGGCTCG TTTAGTGGGG CAGAAAAAAG CGCGTGAAAT TTGGTTCTTA 1025520

TGCCGTCAAT ATAATGCACA AGAAGCATTA GACATGGGCT TAGTCAATAC TGTTGTGCCT 1025580

TACGCTGATC TTGAAAAAGA AACCGTGCGT TGGTGTCGTG AAATGTTACG AAATAGCCCA 1025640

ATCGCGATCC GTTGTTTGAA AGCGGCATTA AATGCGGATT GTGATGGCCA AGCAGGTCTT 1025700

CAAGAATTAG CGGGTAACGC AACAATGTTG TTCTACATGA CTGAAGAAGG TCAAGAAGGC 1025760

CGCAATGCGT TTAACGAAAA ACGCGCCCCA GACTTCAGCA AATTCAGACG TAACCCTTAA 1025820

TTTATCGTTC TAAAGTGCGG TTAAAAATCT TCGTGAATTT TGACCGCACT TTTGTATTGG 1025880

BAD ORIGINAL

DOCID: <WO 9633276A1 1 > BNS page 755

What Is Claimed Is:

1. Computer readable medium having recorded thereon the nucleotide sequence depicted in SEQ ID NO:1, a representative fragment thereof or a nucleotide sequence at least 99.9% identical to the nucleotide sequence depicted in SEQ ID NO:1.

2. Computer readable medium having recorded thereon any one of the fragments of SEQ ID NO:1 depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

3. The computer readable medium of claim 1, wherein said medium is selected from the group consisting of a floppy disc, a hard disc, random access memory (RAM), read only memory (ROM), and CD-ROM.

4. The computer readable medium of claim 3, wherein said medium is selected from the group consisting of a floppy disc, a hard disc, random access memory (RAM), read only memory (ROM), and CD-ROM.

5. A computer-based system for identifying fragments of the *Haemophilus* genome of commercial importance comprising the following elements;

a) a data storage means comprising the nucleotide sequence of SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1;

b) search means for comparing a target sequence to the nucleotide sequence of the data storage means of step (a) to identify homologous sequence(s), and

c) retrieval means for obtaining said homologous sequence(s) of step (b).

BNSDOCID: <WO__9633276A1_I_>

BNS page 1235

6. A method for identifying commercially important nucleic acid fragments of the *Haemophilus* genome comprising the step of comparing a database comprising the nucleotide sequence depicted in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 with a target sequence to obtain a nucleic acid molecule comprised of a complementary nucleotide sequence to said target sequence, wherein said target sequence is not randomly selected.

7. A method for identifying an expression modulating fragment of *Haemophilus* genome comprising the step of comparing a database comprising the nucleotide sequence depicted in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 with a target sequence to obtain a nucleic acid molecule comprised of a complementary nucleotide sequence to said target sequence, wherein said target sequence comprises sequences known to regulate gene expression.

8. An isolated protein-encoding nucleic acid fragment of the *Haemophilus influenzae* Rd genome, wherein said fragment consists of the nucleotide sequence of any one of the fragments of SEQ ID NO:1 depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

9. A vector comprising any one of the fragments of the *Haemophilus influenzae* Rd genome depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

10. An isolated fragment of the *Haemophilus influenzae* Rd genome, wherein said fragment modulates the expression of an operably linked

BNSDOCID: <WO___9633276A1_I_>

open reading frame, wherein said fragment consists of the nucleotide sequence from about 10 to 200 bases in length which is 5′ to any one of the open reading frames depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

11. A vector comprising any one of the fragments of the *Haemophilus influenzae* Rd genome of claim 8.

12. An organism which has been altered to contain any one of the fragments of the *Haemophilus* genome of claim 8.

13. An organism which has been altered to contain any one of the fragments of the *Haemophilus* genome of claim 10.

14. A method for regulating the expression of a nucleic acid molecule comprising the step of covalently attaching 5′ to said nucleic acid molecule a nucleic acid molecule consisting of the nucleotide sequence from about 10 to 100 bases 5′ to any one of the fragments of the *Haemophilus* genome depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

15. An isolated nucleic acid molecule encoding a homolog of any one of the fragment of the *Haemophilus* genome depicted in Table 1a, excluding the fragments of SEQ ID NO:1 depicted in Table 1b wherein said nucleic acid molecule is produced by the steps of:

a) screening a genomic DNA library using any one of the fragments of the *Haemophilus* genome depicted in Table 1a as a target sequence;

b) identifying members of said library which contain sequences which hybridize to said target sequence;

RNS page 1237

c) isolating the nucleic acid molecules from said members identified in step (b).

16. An isolated DNA molecule encoding a homolog of any one of the fragments of the *Haemophilus* genome depicted in Table 1a, excluding the fragments of SEQ ID NO:1 depicted in Table 1b wherein said nucleic acid molecule is produced by the steps of:

a) isolating mRNA, DNA, or cDNA produced from an organism;

b) amplifying nucleic acid molecules whose nucleotide sequence is homologous to amplification primers derived from said fragment of said *Haemophilus* genome to prime said amplification;

c) isolating said amplified sequences produced in step (b).

17. An isolated polypeptide encoded by any one of the fragments of the *Haemophilus influenzae* Rd genome depicted in Table 1a or by a degenerate variant of said fragment, excluding the fragments of SEQ ID NO:1 depicted in Table 1b.

18. An isolated polynucleotide molecule encoding any one of the polypeptides of claim 17.

19. An antibody which selectively binds to any one of the polypeptides of claim 17.

20. A method for producing a polypeptide in a host cell comprising the steps of:

a) incubating a host containing a heterologous nucleic acid molecule whose nucleotide sequence consists of any one of the fragments of the *Haemophilus influenzae* Rd genome depicted in Table 1a or a degenerate

BNSDOCID: <WO___9633276A1_I_>

variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b under conditions where said heterologous nucleic acid molecule is expressed to produce said protein, and

b) isolating said protein.